An interpretation-based account of illusory inferences from disjunction
(draft)

Salvador Mascarenhas, NYU
smasc@nyu.edu

March 15, 2013

1. Introduction

The human capacity for reasoning is subject to failures in remarkably systematic ways, as shown most famously in the Nobel-prize winning work of Daniel Kahneman and Amos Tversky (Tversky and Kahneman, 1974, 1983, to name just two influential papers). Thus far, these phenomena have for the most part been studied by psychologists, who propose accounts rooting failures of reasoning in the general purpose reasoning mechanisms themselves (Tversky and Kahneman, 1974; Johnson-Laird, 1983; Rips, 1994).

In this paper, I outline an alternative source for at least some of our failures of reasoning: an interpretation-based account of some reasoning failures. On an interpretation-based view, "failures of reasoning" is a misnomer, at least for a representative class of attractive but fallacious inference patterns. Rather than being the product of classically unsound general-purpose reasoning mechanisms, some of the "mistakes" we make in fact arise from more complex interpretive processes than meet the eye. Specifically, I show how an important class of reasoning failures, the illusory inferences from disjunction (Johnson-Laird and Savary, 1999; Walsh and Johnson-Laird, 2004), can be accounted for if we assume a classically sound reasoning module that acts upon the pragmatically strengthened meaning of premises, rather than on their literal meaning. Schematically, the illusory inference from disjunction is a conclusion of b from the premises $(a \wedge b) \vee (c \wedge d)$ and a. This inference is fallacious, as it is proved invalid at a world where a, c, and d are true but b is false.

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

The paper is structured as follows. In section 2 I introduce the reasoning data this paper focuses on. I also introduce certain conceptual distinctions important to keep in mind. Section 3 defines the interpretation-based account I propose, based on theories of scalar implicature from formal pragmatics. This theory is put to practice in section 4, where I give a complete account of the illusory inference from disjunction introduced in section 2. Section 5 contains a discussion of the predictions of this account and how in principle they can be tested. Section 6 concludes the paper.

2. FAILURES OF REASONING

This section introduces the reasoning data this paper focuses on, as well as an important distinction between two different classes of reasoning failures.

2.1. Compelling fallacies and repugnant validities

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

(1) P_1: If the card is long then the number is even.
P_2: The number is even.
Conclusion: The card is long.

Under the assumption that P_1 is interpreted as a simple conditional, rather than a biconditional, the inference in (1) is fallacious. It is consistent with both premises that the card not be long, and therefore the conclusion does not follow. Notice that this assumption about interpretation is crucial: if P_1 is interpreted as a biconditional, then the inference does follow. Insofar as we can independently motivate the hypothesis that P_1 is in fact interpreted biconditionally, the inference pattern in (1) lends itself to a simple explanation consonant with an interpretation-based program for failures of reasoning. Unsurprisingly, the fallacy of affirming the consequent has in fact received analyses in the spirit this program. For example, Horn (2000) discusses several historic accounts of the interpretive move from 'if' to 'if and only if' in terms of pragmatic strengthening. Using insights from recent work in formal pragmatics, the account can be simplified as follows.

There is ample evidence for the existence of exhaustification or strengthening operations in natural languages.[1] These operations perform a job very close to what lexical items

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

like English 'only' do.[2] For example, in question answering, it is well established that term answers are most naturally interpreted exhaustively. In what follows, SMALL CAPS indicate focused material.

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

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

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

Suppose now that an exhaustification operation analogous to *only* applies to the conditional premise P_1 of (1), returning its pragmatically strengthened interpretation. The exhaustive interpretation of (4a) would be as in (4b). Conjoining the literal meaning (4a) with the strengthened (4b) then gets us the interpretation in (4c) for the first premise of (1).

(4) a. If the card is long then the number is even.
b. *Only* if the card is long is the number even.
c. If, and only if, the card is long is the number even.

If (4c), rather than (4a), is the interpretation of the first premise of (1), then (1) is in fact classically valid. There is no fallacy, no failure of reasoning.[3]

This account of affirming the consequent is a clear and important precursor to the approach in this paper. Unfortunately, to the best of my knowledge, work from within linguistic semantics on this topic is scarce and has focused entirely on a small number of fallacies discussed in philosophy and rhetoric, mostly related to inferences involving conditionals,

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

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

whose existence scholars have been aware of since classical antiquity.[4] Clearly, a viable interpretation-based program for the study of reasoning failures must extend to the more sophisticated reasoning data discussed today in the psychology of reasoning.

The sizable body of reasoning research conducted by psychologists since the 1960's has produced examples of fallacious inferences far more complex than the example of affirming the consequent in (4). This paper focuses on the illusory inference from disjunction, exemplified in (5), and accepted by approximately 80% of subjects in an experiment by Walsh and Johnson-Laird (2004).

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

The inference in (5) is fallacious: suppose Jane is kneeling by the fire but *not* looking at the window, while Mark is both standing at the window and peering into the garden. This situation would model both premises but falsify the conclusion.

It is useful to abstract away from the specific content given in (5) and consider the schema behind this inference, given in a standard propositional language.[5] The second premise and the conclusion have trivial propositional analyses, but P_1 might be interpreted as an inclusive or an exclusive disjunction. The two options are given as P_1 and P_1' in (6).

(6) P_1: $(a \wedge b) \vee (c \wedge d)$ $\quad\quad$ P_1': $(a \wedge b \wedge \neg(c \wedge d)) \vee (c \wedge d \wedge \neg(a \wedge b))$
P_2: a
Conclusion: b

Notice that either choice of P_1 or P_1' as the interpretation arrived at by subjects in Walsh and Johnson-Laird's (2004) study results in an invalid inference pattern. Both inferences schematized in (6) are falsified at a model making a, c, and d true, but b false. Thus, and unlike the case of affirming the consequent in (5), the illusory inference from disjunction lacks an obvious interpretation-based account. This inference pattern will be the focus of sections 3 and 4, where I show that in fact such an account follows surprisingly from largely

[4]The converse is also largely true: work on fallacies from psychology tends to assume that the interpretive processes are either self-evident or inconsequential. A notable exception is the work of Stenning and van Lambalgen (2008), who combine the methodologies and data of psychology with a deep understanding of non-classical logics and a serious concern for issues of interpretation.

[5]Walsh and Johnson-Laird (2004) took care to ensure that the attractiveness of the pattern in (5) was not merely due to the close internal affinity within the contents of each disjunct—each disjunct being about the same person. Walsh and Johnson-Laird tested every permutation of the atomic (propositional) sentences in P_1 of (5), in particular the variant of (5) where each disjunct of P_1 contains a conjunction whose conjuncts are about different actors. For example: "Either Jane is kneeling by the fire and Mark is standing at the window or otherwise Jane is looking at the TV and Mark is peering into the garden." They also tested examples with four distinct actors. They found a gradual decline in acceptance rates as the number of actors was increased (from only one actor to four distinct actors). However, the difference was only reliable between the one-actor problems (which I do not discuss in this paper) and the four-actor problems.

uncontroversial assumptions about scalar implicature.

Compelling fallacies like the two just exemplified must be distinguished from another kind of failure of reasoning: *repugnant validities*. Repugnant validities are classically valid inference patterns that reasoners often reject. For example, disjunction introduction, exemplified in (7), is notoriously hard to accept (Braine et al., 1984).

(7) P_1: The card is long.
Conclusion: The card is long or the number is even.

Perhaps more surprisingly, *modus tollens* (8) has an appreciably lower acceptability rate than *modus ponens* (Evans et al., 1993; Girotto et al., 1997).

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

A complete interpretation-based account of reasoning failures must either extend naturally to repugnant validities or provide a principled explanation for its inability to account for them. While this paper is exclusively about compelling fallacies, I conclude this section with a remark on the repugnant validity in (7).

The most promising interpretation-based route to explaining compelling fallacies is the phenomenon of scalar implicature, but repugnant validities might require explanations using other elements of pragmatics. For example, resistance to disjunction introduction (7) receives a rather straightforward account purely in terms of primary implicatures (see section 3.2 for a discussion of primary implicatures and their relation to scalar implicatures). By virtue of being a disjunction, the conclusion of (7) prompts an ignorance implicature to the effect that the "speaker" of the sentence is not in a position to assert either of the disjuncts. In particular, the "speaker" is not in a position to assert that the card is long. However, the proposition that the card is long is in fact asserted in the premise of (7). Consequently, the sequence Premise–Conclusion in (7) would be extremely infelicitous in a conversational context, given the primary implicatures of the conclusion.

2.2. A reasoning-based account: mental model theory

The only extant account of human reasoning that predicts the illusory inference from disjunction in (5) is mental model theory (Johnson-Laird, 1983, and a wealth of work by Johnson-Laird and collaborators). While a complete presentation of mental model theory would take us far afield, it is important to see a brief sketch of the account. This will provide a point of departure for the interpretation-based theory.

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

from disjunction, repeated in (9). Since the issue of inclusive or exclusive disjunction is imaterial, I choose the simpler inclusive formulation for ease of exposition.

(9) P_1: $(a \wedge b) \vee (c \wedge d)$
P_2: a
Conclusion: b

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

(10)

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

Next, the reasoning module performs a *conjoin* operation on these premises, combining the mental models. This operation is crucially non-classical. The following is one way to describe it, inspired by the corresponding operation in Koralus and Mascarenhas's (2012) rethinking of mental model theory. Because the model for P_2 is entailed by alternative mental model 1 for P_1, but not alternative mental model 2, reasoners ignore the second alternative model for P_1. Since alternative mental model 2 is no longer being attended to, the conjoined interpretation of the premises consists of only one mental model: a model of a and of b, tacit about everything else. The fallacious conclusion, b, follows immediately from a simple reasoning operation amounting to conjunction elimination.

Mental model theory can account for a large body of data on frequencies of acceptance and rejection of inferences at least as well as its competitors.[6] Because of the dynamic nature of its *conjoin* operation, mental model theory provides immediate accounts of many effects of the ordering of premises on the attractiveness of fallacious conclusions (see for example Girotto et al., 1997). Moreover, mental model theory was the only theory of reasoning that predicted the illusory inferences discussed in this paper (Johnson-Laird and Savary, 1999; Walsh and Johnson-Laird, 2004).

Despite being a major simplification of the mental model account, this sketch illustrates two important problems with the theory. First, mental model theory has been criticized by logicians and philosophers for not being formal enough (Hodges, 1993). This introduces a problem of generality, as the theory is not precise enough to make predictions for novel inference patterns of arbitrary complexity. Second, some elements of the theory lack independent motivation (Koralus and Mascarenhas, 2012). In particular, mental model theory posits linguistically implausible interpretations for some connectives, as well as an operation for combining premises that is largely *ad hoc*.[7]

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

[7] For example, the conditional connective *if*, in mental model theory, is interpreted as a mental model of the

A reasoning-based account like the mental models account just sketched attributes the illusion of validity of the illusory inference from disjunction to the peculiar way in which mental models of the premises are combined. As discussed above, this is by no means the only conceptually plausible theory. An alternative theory would root our failures of reasoning in the interpretive processes that produced representations of the premises in the first place. On this view, "failures of reasoning" is a misnomer: it is not that we make mistakes in reasoning about linguistically presented premises, rather, it is the interpretation of the premises themselves that involves a more complex process than meets the eye.

3.1. Central components of an interpretation-based account

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

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

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

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

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

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

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

The crux of any interpretation-based proposal is ingredient two above: the mechanism for strengthening content. I propose what I take to be the most natural move: to build a theory of reasoning failures based on independently motivated accounts of scalar implicatures. In an *implicature*-based theory, component two above is most naturally concretized as in (11).

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

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

a. assigns to each premise the interpretation required to get the observed reasoning patterns,
b. can provide compelling independent evidence that the interpretations obtained for a. are implicatures in the appropriate sense, and
c. introduces no mischief into the account of implicatures that do not (obviously) bear on matters of reasoning failures.

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

3.2. Sketching an implicature-based account

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

Scalar implicatures are a certain kind of quantity implicatures (both primary and secondary) in which the hearer compares the speaker's utterance S to a certain class of statements that the speaker could have made but chose not to: those statements that result from substituting elements of S with members of their *scales*. The idea that (at least some) lexical items come with lexically stipulated scales was introduced by Horn (1972) to address the *symmetry problem* with Grice's original formulation of the process that calculates quantity implicatures. The traditional view of how scalar implicatures are calculated is as follows. When interpreting an utterance S, a hearer will

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

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

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

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

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

Katzir (2007) and Fox and Katzir (2011) propose to incorporate an appeal to judgments of complexity of the allowed substitutions. The intuition is that, while we want to allow for p to be an alternative to $p \vee q$, we do not want $p \vee q$ to be an alternative to p.[8] On this view, the difference between the two cases is that in the former we consider an alternative no more complex than the original sentence, while in the latter we consider an alternative more complex than the original sentence. If we somehow uniformly prevent more complex substitutions from being considered alternatives, we avoid this problem. First, we must define a relation between syntactic structures, as in (12).

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

Second, I define the substitution source for S as follows.[9]

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

a. the lexicon, and
b. the sub-constituents of S.

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

[9] Fox and Katzir (2011) in fact relativize the notions of complexity and of substitution sources to a context C. This is because they have a third contributor to the substitution source (13), namely the set of salient constituents in a context C. The motivation for this addition comes from data such as (i).

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

Consider the communicative content of the first conjunct: it implicates that it wasn't a little bit more than warm yesterday. On anyone's theory of implicature, this result depends on "It was a little bit more than warm yesterday" being an alternative of the first conjunct. In Fox and Katzir's account of the source of alternatives, this is an instance of substitution of a constituent of the first conjunct ('warm') with a constituent coming not from the lexicon (13a) or the subconstituents of the first conjunct (13b), but in fact from the second conjunct ('a little bit more than warm'), presumably a salient constituent in the context of (i). For the cases I consider in this paper, salient constituents that are not sub-constituents will play no role.

I introduce one final modification of the classical procedure described above. Following Sauerland (2004), I take it that the strengthening procedure from primary implicatures to secondary implicatures must obey the constraint in (14).

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

Formally, (14) corresponds to (15):

(15) Where S is a statement and SA_S is the set of alternatives of S that give rise to primary implicatures, the secondary implicatures of S are the negations of propositions $\varphi \in SA_S$ such that

$$(\forall \psi \in SA_S)\ \neg\varphi \wedge S \not\models \psi\ .$$

3.3. Synthesis

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

1. Compute the alternatives to S that are at most as complex as S (definition in (12)).
2. Collect those alternatives φ that are (1) alternatives to S and (2) strictly stronger than S. Call this set SA_S.
3. Compute primary implicatures: for each $\varphi \in SA_S$, "the speaker does not believe that φ."
4. Compute secondary implicatures: for each $\varphi \in SA_S$ such that the negation of φ does not contradict the literal meaning of S or any of the primary implicatures of S, conclude (that the speaker believes) that φ is false.
5. Call the conjunction of the literal meaning of S together with all of its secondary implicatures the strengthened (exhaustive) meaning of S.

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

3.4. Reasoning with implicatures

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

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

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

(17) **Reasoning in the implicature-based account**
Given a sequence of premises $P_1, \ldots, P_n$ and a conclusion C, begin by calculating the strengthened meaning of each premise, getting the sequence $P_1^+, \ldots, P_n^+$. Then, check to see if the conclusion C follows *classically* from $P_1^+, \ldots, P_n^+$.

The second step of (17) glosses over the issue of what this "checking" procedure consists of. In particular, is it model-theoretic checking, perhaps similar in spirit to mental models accounts, or is it a proof-theoretic process of trying to find a derivation of the conclusion from the premises? As far as I can see, the question is orthogonal to the account of the fallacies given in this paper. It suffices therefore to assume that, however the general-purpose reasoning module works, it targets classically valid reasoning, succeeding if the conclusion follows classically from the premises, and failing otherwise.

One final remark is in order. Since this interpretation-based approach builds on theories of scalar implicature, it is natural to be concerned that the line between reasoning and interpretation has been dangerously blurred. After all, scalar implicatures as described above arise from a process of Gricean *reasoning*. In what sense then is this approach properly *interpretation*-based? The following demarcation line should help. I consider to be reasoning-based any account of fallacies that makes crucial use of non-classical behavior of the *general-purpose* reasoning mechanisms. The mental models approach outlined in section 2.2 falls squarely in this category, given the peculiar way in which the *conjoin* operation is defined. By contrast, I call interpretation-based any theory of fallacies that makes no assumptions about the general-purpose reasoning mechanisms, but rather locates the root of fallacies in language-specific processes. Gricean strengthening procedures as described in this section are language-specific, and can therefore be distinguished from general-purpose reasoning. Accordingly, the predictions of reasoning-based accounts differ from those of interpretation-based accounts, at least in principle. For example, reasoning-based accounts, all things being equal, predict that the same errors will be made with logically equivalent linguistic statements. Interpretation-based accounts make no such prediction, among other reasons, because (otherwise) logically equivalent statements can carry very different implicatures. Predictions will also differ with respect to reasoning about premises not given in

a linguistic form (e.g. reasoning about pictorially given information). I return to this issue in section 5, where I discuss concrete predictions of the interpretation-based account given here.

4. THE ILLUSORY INFERENCE FROM DISJUNCTION

We are now in a position to show how the theory just sketched accounts for the illusory inference from disjunction, introduced in section 2.1. The inference is repeated in (18) and schematized in (19).

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

(19) P_1: $(a \wedge b) \vee (c \wedge d)$
P_2: a
Fallacy: b

To make an in-depth presentation of the account proposed possible, I begin with a slight variant of the schema in (19), with one fewer atomic proposition. This will make the total number of formal alternatives to consider more manageable. As we will see, even this simplification gives rise to a rather rich set of alternatives. The (presumed) fallacious inference pattern I will discuss in detail in this section is the following.

(20) P_1: $(a \wedge b) \vee c$
P_2: a
Fallacy: b

The schema in (20) is still fallacious, and it seems to include everything from (19) that is relevant to the properly studied fallacious inference. Interestingly, the exact schema in P_1 of (20) (but not the more complex P_1 of (19)) was discussed by Spector (2007) and shown to have the same implicature I will derive. The proof I give in this section differs from Spector's in several aspects, but these are largely more expository than substantive.[10]

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

The first step to calculate the strengthened meaning of P_1 is to calculate its set of formal alternatives. It is given in Figure 1 (next page). Each expression in this set of alternatives is

[10] Spector's (2007) objectives were entirely unrelated to mine. He showed a very general correspondence result connecting neo-Gricean approaches to localist accounts of implicature. The few substantive differences between our proofs of the strengthened meaning of P_1 of (20) are due to the fact that Spector had to make some assumptions which I needn't make, given our different purposes.

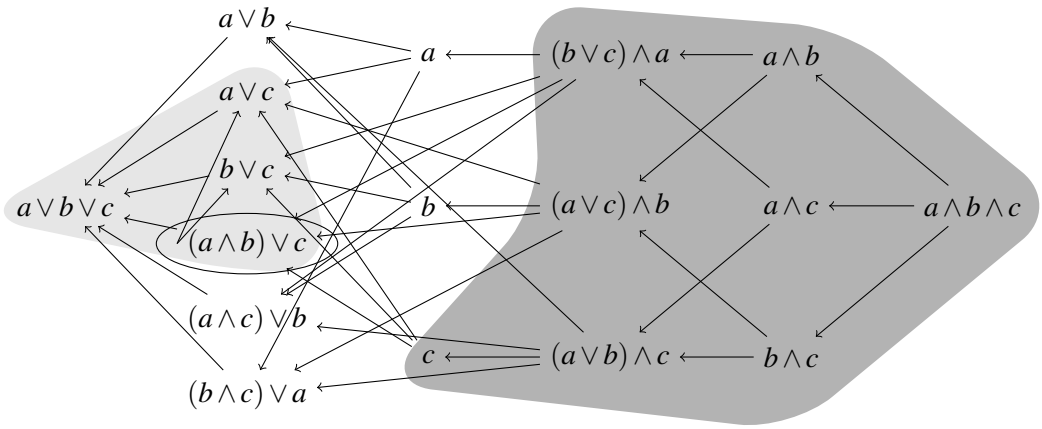

Figure 1: All formal alternatives, up to classical equivalences, for the source $(a \wedge b) \vee c$ (circled in the figure). Arrows between alternatives indicate entailment (transitivity is assumed). The lightly shaded alternatives on the left are weaker than or equivalent to $(a \wedge b) \vee c$. The darker alternatives on the right are strictly stronger than $(a \wedge b) \vee c$.

the result of a licensed substitution according to the adopted theory of formal alternatives. This set is complete, up to certain equivalences we need not worry about.[11]

Next, we calculate primary implicatures for those alternatives that are strictly stronger than P_1.[12] There are eight such alternatives, listed under (21). I indicate already which of

[11]The more technically minded reader will be interested in an informal proof of the completeness of this set. Explaining the procedure I followed will hopefully suffice. First, I considered every substitution of the connectives in the original formula: $(a \wedge b) \vee c$, $(a \vee b) \wedge c$, $a \wedge b \wedge c$, $a \vee b \vee c$. The next step was to consider substitutions and deletions of the propositional atoms for each of these four alternatives. Substitutions of individual atoms for the last two (only $\wedge$s and only $\vee$s), given commutativity and idempotence, will yield formulas equivalent to deletions of atoms, so we can disregard these substitutions. Deletions from these formulas result in all simplex disjunctions and conjunctions possible with the set of atoms $\{a, b, c\}$, as well as each individual atom. All three possible conjunctions and all three possible disjunctions are included in the set of alternatives, as well as all three individual atoms. Consider now the formulas $(a \wedge b) \vee c$ and $(a \vee b) \wedge c$. Deletions from these formulas will result in simplex conjunctions and disjunctions, all of which are already in our list. Substitutions are more interesting in this case. Any substitutions that use an atom more than once (e.g. $(a \wedge b) \vee a$) will be equivalent by absorption to an atom, so we can disregard them. We are therefore left with only two substitution variants for each of $(a \wedge b) \vee c$ and $(a \vee b) \wedge c$, corresponding to "reshufflings" of the propositional atoms. As the reader can verify, all four of these reshufflings are represented in Figure 1.

[12]We could have adopted the view that the relevant alternatives are not just the ones stronger than the literal meaning, but rather all alternatives not weaker than the literal meaning. The set of alternatives giving rise to primary implicatures becomes larger (including unshaded alternatives in Figure 1), therefore predicting a few additional primary implicatures. Interestingly, even with this larger set of primary implicatures, the secondary implicatures will be the same as with the smaller set of primary implicatures gotten by looking only at stronger alternatives. Consequently, my results and analysis of the illusory inference from disjunction are preserved under this alternative view of implicature. I give here a proof sketch. First we need to show that there are no new secondary implicatures. This amounts to proving that each of the negations of the new alternatives, when conjoined with the source of the alternatives, entails some other alternative in Figure 1. The new alternatives are $a \vee b$, $(a \wedge c) \vee b$, $(b \wedge c) \vee a$, a, and b. Each of these, when negated, entails $\neg a$ or $\neg b$, and therefore, conjoined

these alternatives will *also* give rise to secondary implicatures and, for the ones that will not, what the relevant alternatives are that will block their strengthening into secondary implicatures. This will be explained in detail presently.

(21) Alternatives that will give rise to primary implicatures:

Also sec. implic.	Not sec. implic.,	because it would entail
$(a \vee b) \wedge c$	c	$(a \vee c) \wedge b$
$a \wedge c$	$(b \vee c) \wedge a$	c
$b \wedge c$	$(a \vee c) \wedge b$	c
$a \wedge b \wedge c$	$a \wedge b$	c

The predicted primary implicatures are therefore propositions of the form "the speaker is not in a position to assert φ," for each φ in (21).

Now for secondary implicatures. For each of the eight alternatives in (21), we ask whether we can negate that alternative while not contradicting the literal meaning P_1 or any of the primary implicatures. Formally, for S the source of the alternatives and SA_S the set of alternatives stronger than S (the alternatives listed under (21)), we collect all formulas $\varphi \in SA_S$ such that

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

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

(22) For $\textsc{Bot}(\Phi)$ a function returning all elements of Φ that do not asymmetrically entail any other elements of Φ,[13] an alternative $\varphi \in SA_S$ will give rise to a secondary implicature just in case

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

Moreover, notice that, if an alternative φ satisfies the condition in (22), so will any alternatives stronger than φ. This fact will also simplify matters: if we can show that a very

with the source of the alternatives $(a \wedge b) \vee c$, each will entail c, a minimal alternative in Figure 1. Second, we must show that the secondary implicatures we had before will be preserved when we broaden the set of primary implicatures. By (22) and (23), this is equivalent to proving that every element of $\{\{(\neg a \wedge \neg b) \vee \neg c, (a \wedge b) \vee c, \varphi\} : \varphi \in \{a \vee b, (a \wedge c) \vee b, (b \wedge c) \vee a, c\}\}$ is a consistent set of formulas. For $\varphi = c$, the set of formulas is satisfied by a model making c true and both a and b false. For the other cases of φ, the corresponding sets of formulas are satisfied by a model making a and b true and c false.

Since the data discussed so far provide no basis on which to choose between these two theories of strengthening, I opted for the more traditional view of looking only at strictly stronger alternatives. The non-weaker view will be relevant to the discussion of quantified variants of the illusory inference, in section 5.

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

weak alternative will give rise to a secondary implicature, we will also have shown that any alternatives stronger than it will give rise to secondary implicatures.

(23) $(\forall \psi \in \text{BOT}(SA_S))\ \neg\varphi \wedge S \not\models \psi \Rightarrow (\forall \psi \in \text{BOT}(SA_S))\ \neg\varphi' \wedge S \not\models \psi$, for $\varphi' \vDash \varphi$

We are now in a position to see that the table in (21) draws the correct line between the alternatives that give rise to secondary implicatures and those that do not. First, alternative $(a \vee b) \wedge c$ satisfies the condition.[14] It suffices to show that its negation is consistent with the literal meaning $(a \wedge b) \vee c$ and each of the three weakest alternatives, namely c, $(a \vee c) \wedge b$, and $(b \vee c) \wedge a$. In (24), I list the three sets of formulas that must be consistent for $(a \vee b) \wedge c$ to give rise to a secondary implicature, together with a model of each of those sets of formulas, proving their consistency.

(24) The following sets of formulas are consistent

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

By (24) and (23), we establish that the alternatives on the left-hand side of (21) will indeed give rise to secondary implicatures. We must now show that these are *all* the alternatives that will.

Consider c. From $\neg c$ and the literal meaning it would follow that $a \wedge b$ and therefore that $(a \vee c) \wedge b$, which is a member of $\text{BOT}(SA_S)$. Therefore c, by (22), will not give rise to a secondary implicature. A similar reasoning applies to each alternative on the right side of (21): each of these alternatives, if strengthened, would entail the member of $\text{BOT}(SA_S)$ indicated in (21). We are left with the following set of secondary implicatures, corresponding to the negations of the formulas on the left side of (21).

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

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

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

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

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

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

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

which is in turn equivalent to the final strengthened interpretation P_1^+ in (26).

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

The reader can likely already see that (26) will do the required job. Under our simple assumptions about the reasoning component, the (putative) illusory inference in (20) will be judged valid if (27) is classically valid.

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

The pattern in (27) is classically valid.[15] As with the simpler case of affirming the consequent, briefly discussed in section 2.1, there is no fallacy to speak of.

This result extends to the original (unsimplified) illusory inference from disjunction, schematized in (28).

(28) P_1: $(a \wedge b) \vee (c \wedge d)$
P_2: a
Fallacy: b

Simply let c in the demonstration above stand for $c \wedge d$ in P_1 of (28). That the inference in (28) is valid for the strengthened interpretation of its premises in (29) follows as a corollary of the discussion above.

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

However, something even stronger can be shown. The strengthened meaning of P_1 of (28) is in fact the stronger (30).

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

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

Since this result is not strictly needed to derive the original illusory inference from disjunction, I radically abbreviate the proof of this claim.[16] The following two stronger alternatives will be included in the alternatives for P_1 of (28).

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

Both (31a) and (31b) will turn out to satisfy the condition in (22), and will therefore give rise to secondary implicatures. When conjoined with the literal meaning P_1 of (28), the negations of (31a) and (31b) give a formula equivalent to (30).

5. Discussion and empirical predictions

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

But meta-theoretical criteria such as parsimony are not the only tools at our disposal, when trying to decide between a reasoning-based account or an interpretation-based account of compelling fallacies. The interpretation-based account I give in this paper makes a prediction not shared by reasoning-based accounts. If reasoners' acceptance of a certain fallacy hinges on the reasoning module operating on the strengthened meaning of the premises, then acceptance rates should decrease significantly if we manipulate the context, syntactic or pragmatic, in ways that will reliably block the crucial implicatures.

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

I propose a new experimental paradigm to test this prediction of the interpretation-based account. The trick is to convert standard reasoning problems (32) into a conditional form (33).[17]

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

[17] Other reasoning experiments ask "what, if anything, follows" from a sequence of premises, instead of

(32) Standard reasoning problem:
$P_1, \ldots, P_n$
Does C follow from the above premises?

(33) Conditional format:

a. If P_1, then if $\ldots$, then if P_n, then C.
Is the above sentence true?

b. Whenever P_1 and $\ldots$ and P_n, C.
Is the above sentence true?

The conditional schemata in (33) are of course much more syntactically complex than the reasoning problem format in (32). This introduces important concerns about parsing difficulties and ambiguity. For example, naively translating the illusory inference from disjunction stimuli used by Walsh and Johnson-Laird (2004) into conditional format results in ambiguous sentences that will almost certainly be extremely difficult, if not impossible, for subjects to parse in an experiment:

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

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

Consider (35), given in standard reasoning problem format.

(35) P_1: Every boy or every girl is coming to the party.
P_2: John is coming to the party.
Does it follow that Bill is coming to the party?

I find this to be a very compelling inference, and I suspect the reader will agree.[18] It is, however, fallacious: in a model where every girl and John come to the party, and no one else does, the premises are satisfied but the conclusion falsified.

Notice that, in the first premise of (35), the universal quantifiers are equivalent to arbitrarily large conjunctions, assuming that the domain of boys and girls is finite. In a model where b_1 and b_2 are the boys and g_1 and g_2 are the girls, the formulas in (36a) and (36b) are equivalent, and therefore equally apt representations of the literal meaning of P_1 in (35).

(36) a. $((\forall x \in \mathit{boy}')\ \mathit{party}'(x)) \vee ((\forall x \in \mathit{girl}')\ \mathit{party}'(x))$

presenting a putative conclusion C and asking whether C follows. This paradigm can also be converted into conditional form, by giving subjects an incomplete conditional sentence and asking them to complete the *then*-clause in a way that makes the entire sentence true.

[18] Inferences like (35) are part of an ongoing experiment on reasoning that I am conducting jointly with [anonymized]. Pilot tests corroborate the introspection-based intuition that (35) is indeed a compelling fallacy, but reliable results will have to wait for the conclusion of the experiment.

b. $(party'(b_1) \wedge party'(b_2)) \vee (party'(g_1) \wedge party'(g_2))$

The formula in (36b) looks exactly like P_1 of the illusory inference from disjunction discussed in this paper, and similarly for the straightforward representations of P_2 and the proposed conclusion of (35). It is therefore reasonable to call (35) a *quantified variant* of the illusory inference from disjunction.

Moreover, with one modification and one assumption, the theory of implicature I adopt in this paper predicts that (35) should indeed be considered valid by reasoners, since the predicted strengthened meaning for P_1 in (35) will classically validate the inference, in a manner entirely parallel to the propositional illusory inference from disjunction. The required modification is that we consider, rather than only alternatives *strictly stronger* than the literal meaning, alternatives that are simply *not weaker* than the literal meaning, as proposed by Spector (2007).[19] With this expanded set of alternatives, we will get (37b) as an alternative to (37a).

(37) a. Source of the alternatives:
Every boy or every girl is coming to the party.
b. Substituting existential for universal quantifiers and *and* for *or*:
Some boy and some girl are coming to the party.

Strengthening the alternative in (37b) into a secondary implicature, we get the formula in (38).

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

From the conjunction of the secondary implicature in (38) with the literal meaning of P_1 in (35), we get, by distributivity of conjunction over disjunction,

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

Now the required assumption: we must assume that universal quantifiers have existential import, that is, their restrictors are never empty.[20] Consider the first disjunct above. If boys

[19] See footnote 12 on page 13, where I show that the results for the propositional illusory inference from disjunction are preserved under this modification of the theory. For a concise review of the independent arguments in favor of this modification, see also Schlenker (2012).

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

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

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

exist, then that disjunct is equivalent to

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

and similarly for the predicate $girl'$ and the second disjunct. This means that the conjunction of (38) with P_1 in (35) is a formula equivalent to (39a). In (39b) I give an English sentence with the interpretation in (39a), to help parse that formula.

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

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

The reader can easily verify that, with the interpretation in (39a) for P_1 of (35), the proposed inference follows classically, by the same reasoning applied to the original illusory inference from disjunction.

Granting for the purpose of this discussion that (35) is a compelling fallacy, it provides a promising way to test the predictions of the interpretation-based account in this paper. While (34) was ambiguous and very difficult to parse, (40a) and (40b), the conditional formulations of the reasoning problem in (35), are perfectly tractable.

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

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

Recall the prediction of the interpretation-based theory. Since, in (40), the crucial premise of (35) is in the antecedent of a conditional (*if*-clause or *whenever*-clause), the strengthening required to validate the inference is either not predicted to arise or it is predicted to arise less often. Consequently, there should be a significant decrease in acceptability between the standard reasoning problem format in (35) and either of the conditional formats in (40). This provides a general way to falsify any implicature-based account (such as the one given in this paper) of compelling fallacies.[21]

[21] Testing data as in (40) is part of my immediate research plans. It is however important to state a caveat: the reader should be suspicious of any intuitions he or she may have about the truth of (40). Many theories of implicature postulate strengthening operators that can occur in embedded positions in the syntax (e.g., Chierchia, 2004). If these theories are correct, it is not impossible for the relevant *if*-clause in (40a) and *whenever*-clause in (40b) to be interpreted in the stronger way in (39), thereby validating the inference and making the conditional true. Taking this into account, the prediction is not that (40) should never be accepted as true, but rather that (i) there should be a significant difference in the acceptance rates of the reasoning problem vs. the conditional problem, and (ii) this difference in acceptance rates should be correlated with a baseline establishing the rate of implicature calculation for simpler sentences in reasoning problem and in conditional format. Since this is a subtle prediction, the method of introspection is not likely to be reliable, especially in the context of having just processed the same problem given in standard reasoning problem format, as the reader of this paper has. The answer will have to come from a reasoning experiment such as I am currently designing.

6. Conclusion

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

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

But semantics and pragmatics can also benefit from this kind of interaction with the psychology of reasoning. We are interested in the interpretation of linguistic signs, and we study those interpretations partly by inspecting the inferences (entailments, implicatures, presuppositions) validated by utterances. The literature on reasoning should be seen by semanticists as a rich repository of inferences, very many of which should in fact be accounted for by our own theories.

References

Barrouillet, Pierre, Nelly Grosset and Jean-François Lecas (2000). Conditional reasoning by mental models: chronometric and developmental evidence. *Cognition*, 75(3):237–266.

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

Chierchia, Gennaro (2004). Scalar implicatures, polarity phenomena, and the syntax/semantics interface. In A. Belleti, editor, *Structures and Beyond*. Oxford: Oxford University Press.

Evans, Jonathan St B. T., Steve E. Newstead and Ruth M. J. Byrne (1993). *Human reasoning: The psychology of deduction*. Hove, UK: Lawrence Erlbaum Associates Ltd.

Fox, Danny (2007). Free choice disjunction and the theory of scalar implicature. In Uli Sauerland and Penka Stateva, editors, *Presupposition and Implicature in Compositional Semantics*, pages 71–120. Pelgrave McMillan.

Fox, Danny and Roni Katzir (2011). On the characterization of alternatives. *Natural Language Semantics*, 19:87–107.

Girotto, Vittorio, Alberto Mazzocco and Alessandra Tasso (1997). The effect of premise order in conditional reasoning: a test of the mental model theory. *Cognition*, 63:1–28.

Hodges, Wilfrid (1993). The logical content of theories of deduction. *Behavioral and Brain Sciences*, 16(2):353–354.

Horn, Laurence (1972). *On the semantic properties of the logical operators in English*. Ph.D. thesis, UCLA.

Horn, Laurence (2000). From *if* to *iff*: conditional perfection as pragmatic strengthening. *Journal of Pragmatics*, 32:289–326.

Johnson-Laird, Philip N. (1983). *Mental models: towards a cognitive science of language, inference, and consciousness*. Cambridge: Cambridge University Press.

Johnson-Laird, Philip N. and Fabien Savary (1999). Illusory inferences: a novel class of erroneous deductions. *Cognition*, 71:191–229.

Katzir, Roni (2007). Structurally-defined alternatives. *Linguistics and Philosophy*, 30:669–690.

Koralus, Philipp and Salvador Mascarenhas (2012). The erotetic theory of reasoning: formal foundations for the psychology of propositional deductive inferences. Unpublished manuscript: WUSL and NYU.

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

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

Sauerland, Uli (2004). Scalar implicatures in complex sentences. *Linguistics and Philosophy*, 27:367–391.

Schlenker, Philippe (2012). The semantics/pragmatics interface. In Maria Aloni and Paul Dekker, editors, to appear in *Handbook of Semantics*. Cambridge: Cambridge University Press.

Spector, Benjamin (2007). Scalar implicatures: exhaustivity and Gricean reasoning. In Maria Aloni, Paul Dekker and Alastair Butler, editors, *Questions in Dynamic Semantics*. Elsevier.

Stenning, Keith and Michiel van Lambalgen (2008). *Human Reasoning and Cognitive Science*. MIT Press: Cambridge, MA.

Tversky, Amos and Daniel Kahneman (1974). Judgment under uncertainty: heuristics and biases. *Science*, 185:1124–1131.

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

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