Assessing the role of matching bias in reasoning with disjunctions

Mathias Sablé-Meyer
PSL University
NeuroSpin

Salvador Mascarenhas
Ecole Normale Supérieure
Department of Cognitive Studies
Institut Jean-Nicod

Abstract

Mental model theories have been uniquely successful accounting for certain deductive reasoning problems involving disjunctions. In this article we examine an ingredient of most extant accounts of these problems, a *matching* procedure. In two experiments, we show that what is often explained in terms of low-level matching or overlap in content in these theories must in fact take place at a higher level of cognition. We show that no current theory of deductive reasoning offers a satisfactory account of these data, and we outline a version of mental model theory that incorporates insights from Bayesian confirmation theory. We argue that our results point toward a unification of seemingly very disparate kinds of failures of reasoning: our new indirect illusory inferences from disjunction and the conjunction fallacy of Tversky and Kahneman (1983).

1 Introduction

Illusory inferences from disjunction were discovered by Johnson-Laird and Savary (1999) and Walsh and Johnson-Laird (2004). In (1) we show a paradigmatic example of the classical variety of these fallacious inferences.

(1) John speaks English and Mary speaks French, or else Bill speaks German.
John speaks English.
Does it follow that Mary speaks French?
(Adapted from Walsh and Johnson-Laird 2004)

The conclusion does not follow logically: if John speaks English and Bill speaks German, but Mary does not speak French, both premises are satisfied and the conclusion falsified. However, independent studies have shown acceptance rates for the proposed fallacious conclusion around 85% in this and structurally identical problems (Walsh and Johnson-Laird 2004; Mascarenhas and Koralus 2017; Koralus and Mascarenhas 2018).

The pattern in (1) is explained within mental model approaches (Johnson-Laird 1983) with resort to two central elements. First, a special semantics for disjunction, where the

first premise of (1) gives rise to two *alternative mental models*, one for each disjunct. Secondly, a matching procedure: when reasoners notice that the second premise matches part of the first alternative mental model for the first premise, the second alternative mental model drops from attention. The reasoner is left with a model of what remains, *John speaks English and Mary speaks French*, whence the fallacious conclusion follows.

This article investigates the matching component of the general account just sketched. We show that examples such as (2) give rise to illusory inferences from disjunction, but crucially the second premise does not exactly match any element of the first premise. For example, "The trigger was pulled" is certainly *related* to "The gun fired," but it does not *match* it.

(2) The gun fired and the guitar was out of tune, or else someone was in the attic.
The trigger was pulled.
Does it follow that the guitar was out of tune?

We present two experiments establishing the need for a deep revision of the notion of matching. We propose that the solution involves recasting the matching procedure as a *semantic* and *probabilistic* procedure that is sensitive to the content of the material being matched and to causal dependencies between those contents.

From a theoretical standpoint, our results suggest a promising new direction toward bridging the conceptual gap between the study of traditional deductive reasoning problems and seemingly unrelated probabilistic reasoning fallacies. In our concluding remarks, we argue that the conjunction fallacy (Tversky and Kahneman 1983) has the exact same underlying structure as the deductive problem in (2), and we outline a unified theory.

2 Illusory inferences from disjunction and matching bias

The class of illusory inferences from disjunction this article focuses on was discovered by Walsh and Johnson-Laird (2004), from which we show below in (3) a representative example.

(3) Either Jane is kneeling by the fire and she is looking at the TV, or otherwise Mark is standing at the window and he is peering into the garden.
Jane is kneeling by the fire.
Does it follow that she is looking at the TV?

This example and many others in the Walsh and Johnson-Laird (2004) studies have the structure in (4).

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

About 85% of subjects judged that the proposed fallacious conclusion in fact followed.

The materials used by Walsh and Johnson-Laird (2004) are unnecessarily complex to address the question we are interested in. Instead we will use the simpler structure in (5), instantiated in (1).

(5) P_1: $(a \wedge b) \vee c$
P_2: a
Ccl.: b

The structure in (5) generalizes interestingly into a rather diverse paradigm of illusory inferences with *disjunction-like* elements. In particular, these inferences can be reproduced with quantifiers doing the job of conjunction and disjunction (Mascarenhas and Koralus 2017), or with the weak epistemic modal *might* doing the job of disjunction (Mascarenhas and Picat 2019). These results are in line with theories from linguistics on the semantics of indefinite expressions (Kratzer and Shimoyama 2002) and the epistemic modal *might* (Ciardelli, Groenendijk, and Roelofsen 2009), which for entirely independent reasons have proposed that these logical operators have interpretations that share crucial formal properties with disjunction.

2.1 Original mental model theory

Walsh and Johnson-Laird (2004) give the clearest account of illusory inferences from disjunction of the kind we discuss here. This is an account within the original mental model theory, which has been superseded by a revised version in recent years. We discuss how the revised theory fares with respect to the illusory inferences of interest here in the general discussion. Since the revised theory does not have at this point a published discussion of its account of these particular illusory inferences, we focus here on the original mental model theory's account.

Illustrating how the theory gets inferences with disjunction, Walsh and Johnson-Laird (2004) consider exclusivity inferences of the form in (6) below.

(6) *a* or *b* but not both.
a.
Therefore, not *b*

To make this inference, "reasoners can match the categorical information in the second premise with the first of the models [of the disjunction] and then flesh out the model to draw the conclusion *not-b*" (Walsh and Johnson-Laird 2004, 97). This same procedure is meant to account for classical illusory inferences from disjunction.

To the best of our knowledge, the exact workings of the matching procedure haven't been fully spelled out within mental model theory, and the original version of the theory is underspecified in additional ways. In particular it lacks a precise formal regimentation of what mental models are and why disjunctions are interpreted in an idiosyncratic way, giving rise to *sets of alternative* mental models rather than simple mental models.

2.2 Erotetic Theory of Reasoning

Koralus and Mascarenhas (2013) provide a complete formalization of a variant of mental model theory, dubbed the Erotetic Theory of Reasoning (ETR). Incorporating results from linguistic semantics, ETR recasts the mental models account in terms of a question-answer dynamic. ETR builds on the well-established fact that disjunctive sentences share many linguistic properties with questions (Alonso-Ovalle 2006; Groenendijk 2008; Mascarenhas 2009) to propose that reasoners treat the first premise of inferences like (1) as a kind of question: are we in a *John speaks English and Mary French*-situation or are we in a *Bill speaks German*-situation? Reasoners do not like to entertain unanswered questions, so they attempt to find information that will help resolve the question as swiftly as possible. The second premise "John speaks English" *overlaps with* (matches) one of the answers to the question and not the other, so the question is deemed answered in the *John speaks English and Mary French*-direction. Whence it follows that Mary speaks French.

The matching procedure on ETR is given in a fully explicit way, and it requires exact content overlap. This is in line with findings of matching bias elsewhere in the reasoning literature, for example in variants of the Wason selection task (Evans 1999).

Consequently, the erotetic theory of reasoning does not predict a fallacy if the second premise fails to exactly match one of the alternatives provided by the first premise, and is instead merely *related* to it. Consider the schema in (7), of which we gave an example earlier in (2).

(7) $(a \wedge b) \vee c$
d
Does it follow that b?
(where independently d and a are connected)

We investigated indirect illusory inferences of this kind in two behavioral experiments.

3 Experiment 1 — Indirect Illusory Inference from Disjunction

The goal of Experiment 1 was to investigate non-matching relatedness, which we operationalized for simplicity as causal dependence. We hypothesized that the perceived strength of the causal dependence between d and a in the schema in (7) above would have a direct effect on the rate of acceptance of the fallacy.

3.1 Method

This experiment required two disjoint sets of participants, one to *rate* the strength of causal dependencies and the other to perform an inference-making task on patterns like (7).

3.1.1 Participants

Participants were 323 individuals in the United States recruited via Amazon Mechanical Turk. We recruited $160 + 83$ in the rating condition and 80 in the inference condition. All subjects were compensated for participating.

Due to a wording error regarding the reward in the rating condition, we had to cancel our submission as soon as we were notified, correct the wording, and resubmit. Both times the target was 160 participants but we only collected data from 83 in the first submission, which we kept in the analysis, and 156 *new* participants in the second submission. This adds up to a total of 239 participants.

We kept 64 participants in the fallacy experiment: 2 did not correctly report back to the Mechanical Turk website and 14 took the rating experiments prior to this one.

3.1.2 Procedure

Both experiments presented themselves as a web page written in the `jsPsych` library (De Leeuw 2015) with custom plugins developed in our lab. They started with a consent form, followed by instructions, the body of the experiment and then a few demographic questions.

In the rating experiment participants were asked to "indicate the strength of the causal link" for a list of sentences of the form "if [*proposition 1*] then [*proposition 2*]". They were shown 24 conditional sentences, each with a 7-point likert scale ranging from "none" to "perfect." Participants saw three groups of eight conditional sentences, as explained in the Materials section, with repetition of the instructions each time. Two brief pilot experiments were given in between: a brief Stroop task and a single logic question of a very different nature.

The instructions for the inference-making study were to tell whether "a proposed conclusion follows from the sentences." The instructions included an example of a valid inference and an example of an invalid inference, unrelated to the stimuli used in the task, with explanations of why the answer was "yes, the conclusion follows" for one and "no, the conclusion does not follow" for the other. Then participants saw seven illusory inference trials structured as in (7) presented in random order, interleaved with three valid and three invalid controls. For each trial participants could answer yes or no or decide not to answer.

Valid controls were instances of modus ponens whose syntactic complexity was comparable to that of the targets. We used the structure P_1: "If *a* and *b*, then *c*," P_2: "*a* and *b*," "Does it follow that *c*?" Invalid controls followed a similar pattern but the antecedent of the conditional was denied by the second premise. Structurally, P_1: "If *a* and *b*, then *c*," P_2: "not *a*," "Does it follow that *c*?" The role of controls was to establish a baseline for mistakes and get a measure of participants' attention.

3.1.3 Materials

We borrowed the causally connected items (a and d in the schema in (7)) from Cummins (1995), who conducted a study involving causal dependencies much like our rating task.

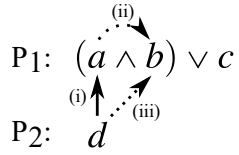

Figure 1: Desired and confounding causal dependencies in our target materials

The rating experiment measured the strength of three kinds of dependencies, schematized in Figure 1. Most importantly, (i) the crucial connection from d to a, which we hypothesized would be predictive of inference-making behavior. We also took two control measures: we checked for (ii) the strength of the connection from a to b, and (iii) that from d to b. This was to make sure that, in the inference-making task, the predicted conclusion b constituted an illusory inference like in the examples in the literature. That is, (iii) if d were to independently lead to b, then a conclusion of b would be explainable purely by the presence of the second premise d. Similarly, (ii) if a independently led to b, and given that d by design was connected to a, the conclusion b would be explained as probabilistic conditional transitivity. Neither of these two scenarios would constitute an illusory inference from disjunction. Accordingly, we decided to keep only those items that showed a moderate or higher connection for (i) d to a, while displaying very weak connections for (ii) a to b and for (iii) d to b.

In (8) is a sample item from each of the three blocks of the rating task just reviewed, and in (9) all of our (i) d to a items.

(8)
i. If the trigger was pulled, then the gun fired
ii. If the gun fired, then the guitar was out of tune
iii. If the trigger was pulled, then the guitar was out of tune

(9)
1. If the brake was depressed, then the car slowed down.
2. If Mary jumped into the swimming pool, then Mary got wet.
3. If the trigger was pulled, then the gun fired.
4. If Larry grasped the glass with his bare hands, then Larry left fingerprints on his glass.
5. If the gong was struck, then the gong sounded.
6. If John studied hard, then John did well on the test.
7. If the apples were ripe, then the apples fell from the tree.

3.2 Analysis and results

3.2.1 Rating

Figure 2 shows the ratings of our 8 item sets in the rating task. We report averages across participants and the standard error. Blocks two and three being controls, we

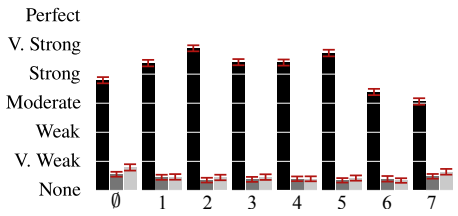

Figure 2: "Strength of the causal connection" rating from the rating experiment. In black is block (i) where participants rated $d \Rightarrow a$, in dark gray is (ii) $a \Rightarrow b$ and in light gray (iii) $d \Rightarrow b$. Standard error is represented in red.

conducted a one way between-subjects ANOVA to compare the effect of the materials on the rating. In block 2 no significant effect was found at the $p < 0.05$ level. In block three there was a significant effect at the $p < 0.05$ level — a post-hoc comparison using the Tukey HSD test indicated that the effect was entirely driven by a single stimulus, the item labeled ∅ in Figure 2, which we therefore removed from the following experiment. We were left with 7 item sets that fulfilled our requirement of variance in the crucial (i) rating but very low ratings for (ii) and (iii).

3.2.2 Inference group

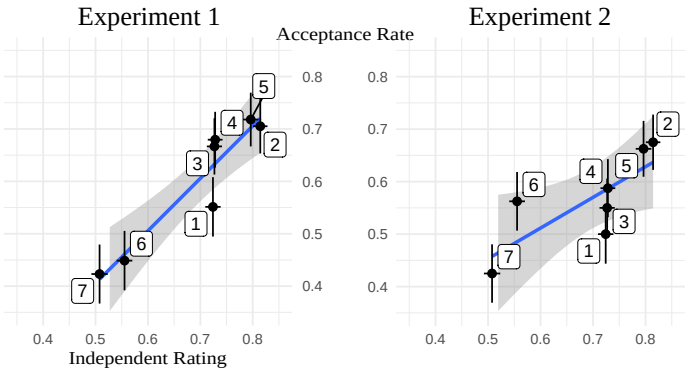

Figure 3: **Top:** Correlation between the average reported strength of the crucial dependence d to a from the rating experiment and the acceptance rate of target fallacies in the inference-making tasks of experiment 1 (left) and experiment 2 (right). We plot each individual item. Horizontal and vertical bars indicate the standard error for both dimensions.

Both valid and invalid controls showed on average around 85% correctness rate (resp. $88.0 \pm 2.1\%$ and $75.2 \pm 2.8\%$). Figure 3, left panel, shows the correlation between the mean acceptance rate of the indirect fallacy and the measured strength of the causal connection.

Acceptance rate in the inference task is significantly predicted by the reported strength

of the entailment in the rating task at the $p < 0.005$ level. Once both variables are scaled to $[0,1]$, the regression has a slope $\beta = 1.00$ ($p = 0.0014$ and $R^2 = 0.89$).

	Estimate	Std. Error	z-value	p-value
Intercept	0.41	0.09	4.59	< 0.001
Rating	4.26	0.84	5.02	< 0.001
Controls	-0.29	0.37	-0.77	0.44
Interaction	13.53	3.55	3.81	< 0.001

Figure 4: Binomial generalized linear model Accepted $\propto$ Rating $*$ Controls after removal of the means for both Rating and Controls.

Additionally, we looked at the way answers to control inferences (valid and invalid) predicted the slope of the correlation between the ratings and the acceptance rate of participants. Table 1 shows the output of a binomial generalized linear model where the interaction term is significant. The effect is as follows: participants who did not succeed on the controls show an absence of effect of the rating on the acceptance rate, and within the others, the higher the score on controls, the tighter the correlation between the rating and the acceptance rate. This suggests that the effect is as present in people as they are engaged in the task, confirming that the fallaciousness of the conclusion does not follow from a lack of attention from participants.

3.3 Discussion

By carefully ruling out internal confounds within each trial, we managed to isolate the main effect that drives the data and measure it separately. Our results show that the extent to which participants accept the fallacious conclusion in the inference-making task is closely positively correlated with an independent measure of the perceived strength of the connection from d to a. Classical illusory inferences where $d = a$, that is the original cases where matching was a plausible account, had an acceptance rate of 85%, suggesting a ceiling effect that matches what we found for valid controls. Given this, it is expected that a more accurate model would be a sigmoid, and that the data samples only over a linear portion of it.

The interaction studies confirm that some participants, 7 in the "low control score," are hardly doing the task and give flat answers throughout. But the data indicates that the more accurate on the controls people are, i.e. the more they are paying attention or the more rational they are, the steeper the slope of the correlation between the rating and the acceptance rate. This indicates that when people are paying attention, they find it easier to resist invalid controls, but still fall for illusory inferences from disjunction, which demonstrates the attractiveness of these fallacies. In other words, the fallacious answer does not solely arise from people generally being irrational, but also from a systematic attractiveness that even participants that were otherwise correct found hard to resist.

The crucial predictor of fallacious behavior is a connection from d to a that relies entirely on world knowledge and cannot be accounted for in terms of a matching algorithm.

4 Experiment 2 — other forms of indirectness

We explored another strategy for inducing the fallacy in an indirect fashion, as schematized in (10) and exemplified in (11).

(10) $(b \wedge d) \vee c$
b
Does it follow that a?
(where independently d and a are somehow linked)

(11) The guitar was out of tune and the brake was depressed, or else someone was in the attic.
The guitar was out of tune.
Does it follow that the car slowed down?

4.1 Method

Experiment 1 showed that the matching part of the mental models and erotetic theory of reasoning accounts cannot be whole story. Experiment 2 investigated whether the sensitivity to world-knowledge causal dependencies was restricted to the interaction between premises, or was operative throughout in these examples. In particular, extant accounts involve matching the second premise of (10) to the first disjunct of the first premise. This leads reasoners to a model of $b \wedge d$ allowing for a conclusion of d by inspection of this model. But do they then conclude that a follows by pursuing the causal dependency from d to a?

The study consisted of a straightforward variant of the inference-making task in experiment 1, where the structure of the fallacious trials was changed from (7) to (10). We recruited 80 participants from the same pool, out of which 10 were already included in a previous experiment and were removed from the analysis. Subjects were compensated for their participation.

4.2 Analysis and Results

Both valid and invalid controls showed on average around 85% accuracy (resp. $90.4 \pm 1.9\%$ and $80.4 \pm 2.5\%$). Figure 3, right panel, shows the correlation between the mean acceptance rate of the indirect fallacy schematized in (10) and the strength of the causal connection as measured in the rating task of experiment 1.

Acceptance rate for target fallacies is significantly predicted by the perceived strength of the crucial connection d to a at the $p < 0.05$ level. Once both variables are scaled to $[0,1]$, the regression has a slope $\beta = 0.58$ ($p = 0.039$ and $R^2 = 0.60$). A binomial generalized linear model to study the interaction yields only the rating of the crucial connection d to a as a good predictor of the behavior at the $p < 0.001$ level.

4.3 Discussion

The significant predictive power of the perceived strength of the connection from d to a over the inference-making behavior shows that contentful, world-knowledge dependencies aren't only operative in the mechanism that combines the information

in the two premises to pick one alternative, but at later steps in the reasoning process. Interestingly, the slope here is less steep, and less of the variance is explained. This suggests that, while these dependencies are operative throughout, the characteristic step of looking for relatedness between the premises in mental model theory and the erotetic theory of reasoning is particularly sensitive to this kind of information. The work we report here allows us to conclude no further than this.

5 General discussion

We've established the existence of indirect illusory inferences from disjunction, where the second premise does not match any model of the first premise, but instead merely raises the probability of a model of the first premise. These illusory inferences have acceptance rates entirely comparable to those of classical illusory inferences from disjunction where matching is a plausible strategy, modulated by the strength of the connection between the non-matching models. We conclude that a more sophisticated process than exact matching is required, one that is sensitive to contentful connections between models. We now turn our attention to various relevant theories of reasoning and assess how they fare with our data.

5.1 Original mental model theory

Classical illusory inferences from disjunction were first discovered and accounted for within the original version of mental models theory (Johnson-Laird and Savary 1999; Walsh and Johnson-Laird 2004). As discussed in the introduction, Walsh and Johnson-Laird (2004) appeal to an underspecified notion of matching to account for these fallacies.

It is possible that by "matching" the authors meant a more sophisticated notion that takes into account world knowledge about causal links and is amenable to modeling varying degrees in the strength of these connections. But such a sophisticated notion would require explication, which is not to be found in the articles discussing these illusory inferences. Taking these articles at face value, the operation is in fact one of direct matching. We conclude that the original mental model theory is either ill equipped to handle our indirect illusions, or the correct account is formulated at too high a level.

5.2 Erotetic theory of reasoning

As discussed in the introduction, the erotetic theory of reasoning of Koralus and Mascarenhas (2013) formalizes the original mental model theory, supplementing it with insights from linguistic semantics that converged with the mental model theory, offering independent motivation for some of its central moving parts. Most importantly, the notion that disjunctive premises, unlike conjunctive ones, give rise to *sets* of mental models converges elegantly with entirely independently motivated theories about the semantic interpretation of disjunctions as sets of *alternative propositions* instead of simple Boolean joins of propositions.

The erotetic theory of reasoning explicitly implements direct matching as the strategy for picking an alternative from the first premise as a function of the information in the second premise. As such, it is ill suited to account for the indirect inferences discovered in this article.

5.3 Revised mental model theory

The original mental model theory of Johnson-Laird and collaborators underwent a major revision in the recent past, most clearly presented by Khemlani, Byrne, and Johnson-Laird (2018). For our purpose of assessing the new mental model theory account of illusory inferences from disjunction, the following innovations are of central interest.

1. The division of labor between intuitive (System 1) processes and deliberate (System 2) processes has been revised. System 2 works with fully explicit models, on which more below. System 1 works with underspecified mental models that do not include explicit negations of mental models not asserted in the premises, manifesting what used to be called the *principle of truth*. Importantly, a novel parameter γ determines the degree of tolerance of the notion of *necessary conclusion*: System 1 operating with low γ will consider as *weak necessities* some cases that with high γ are considered merely *possible* conclusions.[1]

2. The theory makes explicit use of a *modulation* process that takes world knowledge into account in the interpretation of mental model premises.

We find much to commend in the revised version of the theory. In particular, the precise formulation of modulation seems well-suited to account for the reliance on world knowledge in the new data we present in this article. In other words, insofar as the revised mental model theory can account for *classical* illusory inferences from disjunction, where matching was a possibility, modulation will take care of extending that account to an account of the new *indirect* inferences we investigate in this article.[2]

However, the revised mental model theory's coverage of *classical* illusory inferences from disjunction is problematic. We see two ways in which the theory might incorporate classical illusory inferences from disjunction, but each strategy comes with its own issues and open questions. We discuss these strategies and our reservations about them presently.

[1] Khemlani, Byrne, and Johnson-Laird (2018) describe γ as the probability that reasoning relies on the notion of *weak necessity* to draw conclusions. This suggests that a high value of γ should prompt reliance on weak necessity. But the lisp implementation of the theory `mSentential` in facts works the other way around: the lower γ is the higher the rate of weak-necessity conclusions. We use the scale for γ defined in `mSentential`.

[2] One caveat that should be mentioned is that it is not entirely clear how the revised mental model theory accounts for the tight linear relation between the rate of acceptance of the fallacies and the judged strength of the relevant causal connections. Mental model theory has an elegant and parsimonious way of formalizing probabilistic reasoning that does not use explicit probability measures. Instead, the theory postulates a process that considers the proportion of models in a mental model set that support a certain conclusion. But the indirect illusions in this article rely on world-knowledge connections that do not come with explicit alternatives, so that it is unclear how to represent these connections and their impact on reasoning without explicit probability measures. We outline a positive proposal using probabilities in the conclusions to this article.

5.3.1 System 2 processes under low γ

If the parameter γ is low enough, then *weakly necessary conclusions* will be drawn. For Γ the models of the premises and φ those of the conclusion, φ will be a weak necessity of Γ just in case (i) every model of φ is supported by the premise models Γ, and (ii) there is some model of the premises that does not support any model of the conclusion. Support in turn is defined as follows. A model m' supports a model m just in case m is included in m'. This notion generalizes to sets of mental models: a set of premise models Γ supports a model m just in case there is some model m' in the premises such that m' supports m.

In less formal terms, φ will be a weak necessity of Γ just in case (i) every model of the conclusion is included in some model of the premises, and (ii) there are orthogonal models of the premises, that is some models of the premises do not contain any models of the conclusion. Classical illusory inferences from disjunction will come out as weak necessities under this definition. Recall the structure of classical illusory inferences from disjunction in (5), repeated below as (12).

(12) P_1: $(a \wedge b) \vee c$
P_2: a
Ccl.: b

The conjunction of the premises in classical illusory inferences from disjunction has two mental models: $ab\not{c}$ and $a\not{b}c$ (a crossed out letter such as $\not{x}$ represents the explicit negation of x). Considering the conclusion b, it is clear that indeed (i) every model of the conclusion (there is only one, namely b) is included in a model of the premises (in this case $ab\not{c}$), but (ii) there is a model of the premises ($a\not{b}c$) that does not include any model of the conclusion. This provides a System 2 account of these illusory inferences under low γ. It is not entirely clear whether this is desired, for System 2 is about deliberate reasoning with fully explicit models and should be resistant to fallacious reasoning. However, we submit that in and of itself this does no harm, since system 2 under high γ does resist the fallacy. System 2 under high γ implements a stronger notion of necessity that requires that *every* model of the premises support the conclusion.

There is an important wrinkle in this System 2 account of classical illusory inferences from disjunction. The proof we just sketched can be immediately adapted into a proof that the schema in (12) supports a conclusion of c as a weak necessity. The conclusion model c is included in model $a\not{b}c$ of the premises, and there is still a model of the premises (now $ab\not{c}$) that does not include the model of the conclusion.

This prediction is problematic, for the attractiveness of the two patterns is sharply different. Consider:

(13) John speaks English and Mary speaks French, or else Bill speaks German.
John speaks English.
Concl.: Mary speaks French.

(14) John speaks English and Mary speaks French, or else Bill speaks German.
John speaks English.
Concl.: Bill speaks German.

The example in (13) is a well-known illusory inference from disjunction, with acceptance rates in the order of 85%, while (14) is, we submit, either *not at all* a compelling fallacy, or it is a very weak illusion, by no means comparable to the high acceptance rate of (13). Further experimental work is required, but it seems clear to us that these two reasoning patterns are sharply different, and that the System 2 account under low γ just sketched altogether lacks the ability to make this distinction.[3]

5.3.2 System 1 processes — a pragmatic confound

System 1 works with underspecified models of the premises, which represent only what is explicitly asserted. For the classical illusory inference schematized in (12), the models of the first premise $(a \wedge b) \vee c$ are as in (15). Notice in particular the gaps: the first model does not represent the negation of c, nor does the second model represent the negations of a and b.

(15) $a\ b$
$\quad\ c$

The second premise of the classical illusory inferences has only one simple model: a. Following the theory as presented by Khemlani, Byrne, and Johnson-Laird (2018), the next step is to conjoin the models of these two premises. One asks, for each model of the first premise in (15), whether the conjunction with the second premise a yields a consistent or an inconsistent model. If the result is consistent, the new model is part of the resulting conjunction, otherwise it is not. The crucial question is the following: does the second line in (16) propose a consistent or an inconsistent combined model? What is the outcome of this procedure? The answer is unclear.

(16) $a\ b \quad + a$
$\quad\ c + a$

In the version of the theory in print, two models are inconsistent when "one represents a proposition and the other its negation" (Khemlani, Byrne, and Johnson-Laird 2018, 9–10). The left-hand side of the second line of (16) does *not* represent the negation of a. This is one of the defining characteristics of mental models, they are underspecified and represent only what is true, not what it false. Consequently, the resulting models after conjunction are ab and ac from which both the b and c conclusions are weak necessities. This is much like the situation discussed in the previous section for System 2 processes under low γ, and therefore subject to the same concerns.

However, these predictions conflict with the behavior of `mSentential`, the lisp implementation of the theory developed by the authors. On `mSentential`, we observe that

[3]The original mental model theory was in agreement with our judgments here. As explained by Walsh and Johnson-Laird (2004), not only is a *c* conclusion *not* predicted for examples as in (14), in fact what is predicted on that theory is a conclusion of *not-c*.

the fallacy is derived as a *strong* necessity, that is under high γ, and that the c conclusion is never acceptable.

Inspecting the lisp implementation reveals that it builds the two mental models we expect for the first premise and represents only what is true as in (15). However, when the second premise, a, is conjoined with the models in (15), explicit negations are added by `mSentential` and only one model is present, namely $ab\not{c}$. From here the fallacious conclusion b (strongly) follows immediately, while the unobserved conclusion c does *not* follow.

This is a promising result, but it relies on `mSentential` behavior that is surprising in view of the presentation of the theory that exists in print (Khemlani, Byrne, and Johnson-Laird 2018), which we just reviewed. In a nutshell, it is puzzling that conjunction with the second premise a should eliminate the second model of (15).

As far as we can tell, the implementation in `mSentential` uses "flags" of a sort, which keep track of relevant propositional letters that are not represented. These are c for the first line of (15), and a and b for the second line of (15). These flags will be taken into consideration when conjunctions are computed. In virtue of conjunction, the models effectively act as if explicit negations were being represented for each relevant proposition not asserted by the model. In other words, the moment mental model conjunction occurs, the mental models for the first premise of our illusory inferences from disjunction are not as in (15), but are for all relevant purposes indistinguishable from the much stronger ones in (17).

(17) $ab\not{c}$
$\not{a}\not{b}c$

The way in which these models are stronger is not without interest. The models in (17) for the first premise of the illusory inference from disjunction correspond in fact to what formal pragmatics calls a strongly exhaustive interpretation. These interpretations were first discussed by Spector (2007) in an entirely independent context, they were used by Mascarenhas (2014) to argue for the possibility of an absolving interpretation for these fallacies, and they were carefully investigated by Picat (2019), who gave experimental evidence that they were scalar implicatures resulting from pragmatic processes, but that they did not explain the totality of the phenomenon of illusory inferences.

It is interesting to contrast this account with that of (our reading of) the classical mental model theory, or the erotetic theory. These accounts used a notion of matching that derived the fallacy in terms of effects of attention, modeled as a question under discussion on the erotetic variant of the mental model theory. On these accounts, the second premise a matched part of the first model of the first premise in (15), and consequently the second model of (15) dropped from attention. By contrast, on `mSentential` the second model is more than merely forgotten, it is considered to be directly inconsistent with the second premise.

To sum up, we see two possible interpretations of the revised mental model theory in the case of System 1. If one follows the presentation by Khemlani, Byrne, and Johnson-Laird (2018), the fallacious conclusion can be derived, but to the same extent

as the unobserved conclusion c. If one follows the implementation in `mSentential`, the fallacious conclusion can be derived, but the mechanism whereby this is done is indistinguishable from a pragmatic mechanism of scalar implicature. This is not necessarily an unwanted result, but it raises at least two important questions. First, one would want to understand in what way mental models do not represent what is false, if indeed the moment conjunction with a second premise is involved they function *as if* they represented the negation of every relevant proposition that is not explicitly affirmed. Second, the broader picture of illusory inferences becomes quite mysterious. In particular, illusory inferences with indefinites (Mascarenhas and Koralus 2017) and illusory inferences with epistemics (Mascarenhas and Picat 2019) cannot be accounted for in terms of pragmatics.

5.4 Probabilistic theories — the new paradigm

There are very successful ways to account for reasoning under uncertainty with the probability calculus (see in particular Oaksford and Chater 2007, 2007; Adams 1996; Johnson-Laird, Khemlani, and Goodwin 2015), modeling subjective probabilities. These theories have been particularly insightful on the broad and important topic of reasoning with conditionals. This literature is however sparser on the topic of reasoning with alternatives, such as provided by disjunction and disjunction-like elements.

As far as we can see, this family of theories is not yet well equipped for the kind of problem we discuss in this article. In particular they have same issue as the one just discussed for the revised mental model theory in distinguishing the b conclusion from the c conclusion under flat equiprobable priors, that is $P(a) = P(b) = P(c) = \frac{1}{2}$.

There are at least two ways of operationalizing subjective validity for such theories. One of them is p-validity, under which a conclusion follows to the extent that it is no less probable than its premises. This will fail to separate the b conclusion from the c conclusion because $P((a \wedge b) \vee c, a) = P((a \wedge b) \vee (a \wedge c)) = \frac{3}{8}$, and since the priors are $P(b) = P(c) = \frac{1}{2}$ both conclusions b and c end up equally p-valid.

An alternative is to compare not the prior probabilities but the posterior probabilities of the putative conclusions on the premises. Once again this cannot distinguish b from c at least under the assumption of independence and flatness of priors, for $P(b|(a \wedge b) \vee c, a) = \frac{2}{3} = P(c|(a \wedge b) \vee c, a)$, and therefore both conclusions b and c should be equally acceptable.

6 Conclusions and ongoing work

We have shown that previous accounts of illusory inferences from disjunction posited matching algorithms where in fact a much more sophisticated process was operative. This process is sensitive to dependencies between propositions that recruit world knowledge, as demonstrated by the close correlation between assessments of the strength of the dependence and rates of commission of the target fallacy. We further concluded that other powerful and insightful approaches to reasoning, in particular the revised

mental model theory and probabilistic approaches, are ill-equipped to deal even with the *classical* examples of illusory inferences from disjunction.

These observations matter. When they were discovered, illusory inferences from disjunction weren't necessarily thought to be more than just another data point to add to our catalog of failures of deductive reasoning. But recent work at the intersection of reasoning and linguistic semantics has shown that these illusory inferences are the tip of a much larger and more interesting iceberg, which can be informally but usefully characterized as reasoning with *alternatives* that prompt question-answer dynamics. Disjunctions are the generators of question-like alternatives *par excellence*, but they are by no means the only ones. So far the literature has identified indefinites and weak modal operators as inducers of illusory inferences that superficially seem entirely unrelated to the original illusory inferences from disjunction. Consequently, understanding just how the alternatives prompted by the first premise (the question) are manipulated by attempts to match them with the second premise (the answer) is an important step toward understanding human reasoning with alternatives.

Additionally, the indirect illusory inferences from disjunction in this article may play a very useful role connecting the study of failures of deductive reasoning to the study of better known, and perhaps more ecologically valid problems. In the most famous example of the conjunction fallacy (Tversky and Kahneman 1983), a description d of an individual suggests that she is interested in social activism. Subjects are given two options: "bank teller" and "bank teller who is active in the feminist movement." Overwhelmingly they consider that the conjunctive option is most likely to be true. Structurally, this is very much like our indirect illusory inferences from disjunction:

(18)

Indirect illusory inference	**Conjunction fallacy**
$(a \wedge b) \vee c$	$(b \wedge f) \vee b$
d, which points to a	d, which points to f
Conclusion: $a \wedge b$ (whence b)	Conclusion: $b \wedge f$

As (18) shows, the conjunction fallacy can be seen as a more complex special case of our indirect illusory inferences from disjunction. Studying indirect illusory inferences can thus in principle be revealing of the reasoning processes behind the conjunction fallacy, for indirect illusory inferences involve the same structure while removing a number of extraneous elements from the usual materials used in conjunction fallacy experiments. In particular, indirect illusory inferences from disjunction do not rely on individuating information or stereotypes.

In ongoing work, we are formalizing and testing a Bayesian confirmation-theoretic implementation of the erotetic theory of reasoning that extends to the data we report in this article and cashes out the connection with the conjunction fallacy. We take the two alternatives provided by the first premise to propose competing *hypotheses*, and the categorical information in the second premise to provide adjudicating *evidence*. We propose that reasoners compare the extent to which the evidence *increases the firmness* of each hypothesis, and choose the hypothesis that maximizes this measure. This is a contrastive version of Bayesian confirmation theory (Fitelson 2013), which has

been shown to explain behavior in the conjunction fallacy (Crupi, Fitelson, and Tentori 2008). For the schematic indirect illusory inference in (18), reasoners are considering two competing hypotheses provided by the first premise: $a \wedge b$ and c. The evidence in the second premise, d, raises the probability of the first hypothesis and not the second, thereby increasing its firmness. This process is of necessity a content-sensitive mechanism, and thus it can handle in the same way classical illusory inferences from disjunction, the new indirect variants, and the conjunction fallacy.

7 Acknowledgments

The authors would like to thank Benjamin Spector, Emmanuel Chemla, Emile Enguehard, Philipp Koralus, Ruth Byrne, Vincenzo Crupi, Philippe Schlenker, and three anonymous reviewers for Cog Sci 2019 for very helpful feedback.

This work was supported by Agence Nationale de Recherche under grant ANR-17-EURE-0017 (FrontCog, Department of Cognitive Studies, ENS); and Agence Nationale de Recherche under grant ANR-18-CE28-0008 (LANG-REASON). No potential conflict of interest was reported by the authors. The data that support the findings of this study are available from the corresponding author, M. Sablé-Meyer, upon request. As part of a biger lab project, open sourcing the data is planned.

References

Adams, Ernest W. 1996. "A Primer of Probability Logic."

Alonso-Ovalle, Luis. 2006. "Disjunction in Alternative Semantics." PhD diss., UMass Amherst.

Ciardelli, Ivano, Jeroen Groenendijk, and Floris Roelofsen. 2009. "Attention! Might in Inquisitive Semantics." In *Proceedings of the 19th Conference on Semantics and Linguistic Theory (Salt)*, 91–108.

Crupi, Vincenzo, Branden Fitelson, and Katya Tentori. 2008. "Probability, Confirmation, and the Conjunction Fallacy." *Thinking & Reasoning* 14 (2): 182–99.

Cummins, Denise D. 1995. "Naive Theories and Causal Deduction." *Memory and Cognition* 23 (5): 646–58.

De Leeuw, Joshua R. 2015. "JsPsych: A Javascript Library for Creating Behavioral Experiments in a Web Browser." *Behavior Research Methods* 47 (1): 1–12.

Evans, Jonathan St B. T. 1999. "The Influence of Linguistic Form on Reasoning: The Case of Matching Bias." *The Quarterly Journal of Experimental Psychology* 52 (1): 185–216.

Fitelson, Branden. 2013. "Contrastive Bayesiansim." In *Contrastivism in Philosophy*, edited by Martijn Blaauw. Routledge.

Groenendijk, Jeroen. 2008. "Inquisitive Semantics: Two Possibilities for Disjunction." In *Proceedings of the Seventh International Tbilisi Symposium on Language, Logic and Computation*.

Johnson-Laird, Philip N. 1983. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cambridge: Cambridge University Press.

Johnson-Laird, Philip N., Sangeet Khemlani, and Geoffrey P. Goodwin. 2015. "Logic, Probability, and Human Reasoning." *Trends in Cognitive Sciences* 19 (4): 201–14.

Johnson-Laird, Philip N., and Fabien Savary. 1999. "Illusory Inferences: A Novel Class of Erroneous Deductions." *Cognition* 71 (3): 191–229.

Khemlani, Sangeet, Ruth M. J. Byrne, and Philip N. Johnson-Laird. 2018. "Facts and Possibilities: A Model-Based Theory of Sentential Reasoning." *Cognitive Science* 42 (6): 1887–1924.

Koralus, Philipp, and Salvador Mascarenhas. 2013. "The Erotetic Theory of Reasoning: Bridges Between Formal Semantics and the Psychology of Deductive Inference." *Philosophical Perspectives* 27: 312–65.

———. 2018. "Illusory Inferences in a Question-Based Theory of Reasoning." In *Pragmatics, Truth, and Underspecification: Towards an Atlas of Meaning*, edited by Ken Turner and Laurence Horn, 34:300–322. Current Research in the Semantics/Pragmatics Interface. Leiden: Brill.

Kratzer, Angelika, and Junko Shimoyama. 2002. "Indeterminate Pronouns: The View from Japanese." In *Third Tokyo Conference on Psycholinguistics*.

Mascarenhas, Salvador. 2009. "Inquisitive Semantics and Logic." Master's thesis, ILLC.

———. 2014. "Formal Semantics and the Psychology of Reasoning: Building New Bridges and Investigating Interactions." PhD thesis, New York University.

Mascarenhas, Salvador, and Philipp Koralus. 2017. "Illusory Inferences with Quantifiers." *Thinking and Reasoning* 23 (1): 33–48.

Mascarenhas, Salvador, and Léo Picat. 2019. "*Might* as a Generator of Alternatives: The View from Reasoning." In *Proceedings of Salt 29, Ucla*.

Oaksford, Michael, and Nicholas Chater. 2007. *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*. Oxford University Press.

Picat, Léo. 2019. "Inferences with Disjunction, Interpretation or Reasoning?" MA thesis (CogMaster), Ecole Normale Supérieure. http://web-risc.ens.fr/~lpicat/website/picat-m2-thesis-pre-print.pdf.

Spector, Benjamin. 2007. "Scalar Implicatures: Exhaustivity and Gricean Reasoning." In *Questions in Dynamic Semantics*, edited by Maria Aloni, Paul Dekker, and Alastair Butler. Elsevier.

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.