Assessing the role of matching bias in reasoning with disjunctions

(article under review)

Mathias Sablé-Meyer
NeuroSpin CEA, Cognitive Neuroimaging Unit INSERM U992, Saclay, France
PSL University, Collège de France, Paris, France

Salvador Mascarenhas
Institut Jean Nicod, Département d'études cognitives
Ecole Normale Supérieure
ENS, EHESS, PSL University, CNRS, Paris France

Abstract

On mental models theories, reasoners create mental representations of information which they manipulate in order to derive new conclusions. These theories have been uniquely successful at explaining a class of attractive fallacies involving disjunctions. In this article we examine a crucial ingredient of mental models accounts of these illusions, a *matching* procedure. In three experiments, we show that what is 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 introduce variants of illusory inferences from disjunction whose acceptance by participants is accurately predicted by their confidence in causal connections that rely on world-knowledge.

Keywords: mental models; reasoning; fallacies; disjunctions; matching bias;

Introduction

There is a paradigm of compelling fallacies involving disjunction, of a kind unfamiliar to most philosophers and psychologists of reasoning, called illusory inferences from disjunction.

(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 & Johnson-Laird, 2004)

The conclusion does not follow classically: 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 & Johnson-Laird, 2004; Koralus & Mascarenhas, 2016; Mascarenhas & Koralus, 2017).

The pattern in (1) is explained within mental models approaches (P. N. 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 by an analog of conjunction elimination.

This article investigates the matching component of the general account just sketched. We show that examples such as (2) and (3) 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?

(3) This party is only for French people and linguists.
That guy is European.
Does it follow that he is French?

We present three 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* procedure that is sensitive to the content of the material being matched and to causal dependencies between those contents.

Illusory inferences from disjunction and syntactic matching

The class of illusory inferences from disjunction this article is about was discovered by Walsh & Johnson-Laird (2004), from which we show below in (4) a representative example.

(4) 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 all others in the Walsh & Johnson-Laird (2004) study have the structure in (5), where '$\wedge$' stands for "and" and '$\vee$ for "or."

(5) 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 & Johnson-Laird (2004) are

unnecessarily complex to address the question we are interested in. Instead we will use the simpler structure in (6), instantiated in (1) on page 1.

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

Mental models accounts

Mental models theory of reasoning and its close relatives offer the best extant accounts of these illusory inferences.[1] In a nutshell, the mental models account runs as follows.

(i) Reasoners build minimal mental representations, or mental models, that verify each of the premises.

(ii) Disjunctive premises are represented as sets of alternative mental models.

(iii) P_1 in (6) gives rise to a set of two alternative models: a minimal model of $a \wedge b$ and a minimal model of c.

(iv) Upon hearing P_2 (a), reasoners notice that it *matches* the first alternative model for P_1 ($a \wedge b$), but not the second (c). The second model thus drops from attention.

(v) The combined representation of the premises is therefore only one mental model: $a \wedge b$. From here, b follows by conjunction elimination.

To the best of our knowledge, the exact notion of *matching* employed in step (iv) for illusory inferences from disjunction is never fully spelled out. P. Johnson-Laird & Savary (1999) however give two tables of principles from which it is clear that the relevant intended notion is one of overlap in atomic propositions. For example, a mental model representing two atomic propositions a and b will match a mental model with a. This much will work as an account of classical illusory inferences as in (1). We will show shortly that this notion of matching fails for a new class of illusory inferences.

The original description of mental models is underspecified in some ways. In particular it lacks a precise formal regimentation of what mental models are, what the mechanisms of the matching procedure are, and why disjunctions are interpreted in an idiosyncratic way, giving rise to *sets of alternative* mental models rather than simply mental models.

Erotetic Theory of Reasoning Koralus & Mascarenhas (2013) provide a complete formalization of a variant of mental models theory, dubbed the Erotetic Theory of Reasoning (ETR). Incorporating results from linguistic semantics, ETR recasts the mental models account just reviewed 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, among many others) to propose that reasoners treat the first premise of inferences like (6) 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, but it still requires content overlap, much like classical mental models theory. Consequently, neither account predicts 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 example in (2), on the preceding page, schematized in (7).

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

We investigated the schema in (7) in behavioral experiments.

Experiment 1 — Indirect Illusory Inference from Disjunction

The goal of Experiment 1 was to investigate non-matching relatedness, which we operationalized 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.

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).

Participants Participants were 240 individuals in the United States recruited via Amazon Mechanical Turk. We recruited 160 in the rating condition and 80 in the inference condition.

For the analysis we kept 153 participants in the rating experiment: 2 did not correctly report back to the Mechanical Turk website and 5 took an earlier pilot experiment that also leveraged on illusory inferences from disjunction. We kept 64 participants in the fallacy experiment: 2 did not correctly report back to the Mechanical Turk website and 14 took both experiments. Subjects were compensated for participating.

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

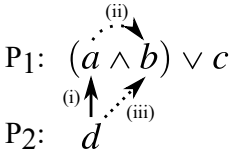
[1] For example, the probability heuristic model of reasoning (Oaksford & Chater, 2007) has a hard time accounting for these fallacies. This is because $P(b|(a \wedge b) \vee c, a) = 1/3 = P(c|(a \wedge b) \vee c, a)$, that is the theory would predict that subjects should find b and c equally acceptable, which is certainly not the case. See Mascarenhas & Koralus (2017) for a more detailed discussion.

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 our target and control inferences, 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) interleaved with three valid and three invalid controls, for a total of 13 items presented in random order. For each trial participants could answer yes or no or decide not to answer.

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.



The rating experiment measured the strength of three kinds of dependencies, schematized in the figure to the left. 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. Additionally, (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

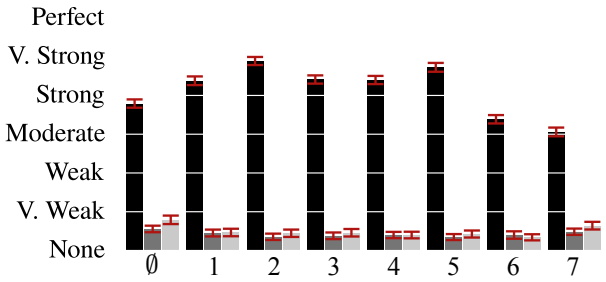


Figure 1: "Strength of the causal connection" rating from the rating experiment. In black is block (i) where participants rated $d \Rightarrow a$, in dark gray is (ii) $a \Rightarrow b$ and in light gray (iii) $d \Rightarrow b$. Standard error is represented in red.

(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.

Analysis and results

Rating Figure 1 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 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 $\varnothing$ in figure 1, 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).

Inference group Both valid and invalid controls showed on average around 85% correctness rate (resp. $87.5 \pm 2.4\%$ and $75.6 \pm 3.2\%$). Top of figure 2, 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 $\beta = 0.97$ ($F[1,5] = 45.07, p = 0.0011$) with $R^2 = 0.90$.

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. Figure 2 shows the slope for three groups and Table 1 shows the output of a binomial generalised linear model.

	Estimate	Std. Error	z-value	p-value
Intercept	0.50	0.10	4.98	< 0.001
Rating	4.14	0.93	4.42	< 0.001
Controls	-0.37	0.41	-0.91	0.36
Interaction	13.0	3.89	3.35	< 0.001

Table 1: Binomial generalised linear model Accepted$\propto$Rating $*$ Controls after removal of the means for both "rating" and "controls".

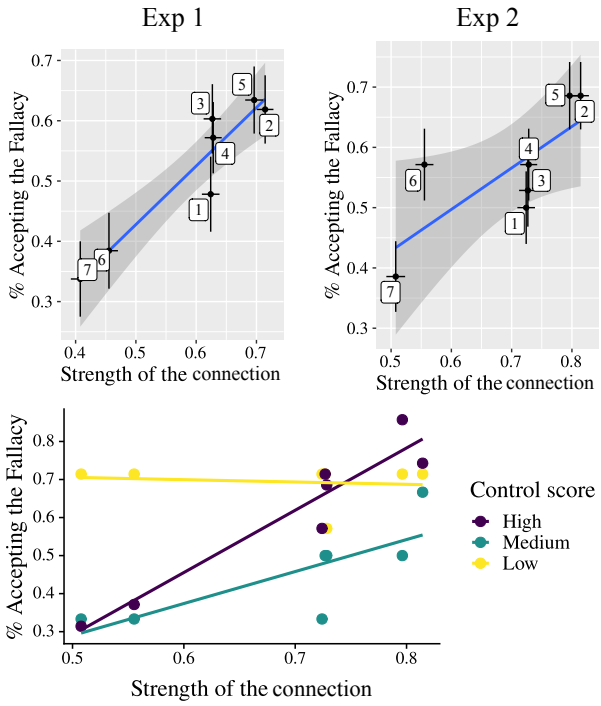


Figure 2: **Top:** Correlation between the average reported strength of the crucial dependence d to a from the rating experiment and the acceptance rate of target fallacies in the inference-making tasks of experiment 1 (left) and experiment 2 (right). We plot each individual item. Horizontal and vertical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

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. This suggests that a better model might be an S-curve.

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 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. This demonstrates the attractiveness of these fallacies.

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.

Experiment 2 — other forms of indirectness

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

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

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, step (v) of the mental models account involves a step of conjunction elimination. Considering the schema in (10), we are led to a model of $b \wedge d$ and can get d by conjunction elimination, but do we 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.

Analysis and Results

Both valid and invalid controls showed on average around 85% accuracy (resp. $89.5 \pm 2.1\%$ and $80.0 \pm 3.3\%$). Top of figure 2, 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 $\beta = 0.69$ ($F[1,5] = 6.93, p = 0.046$) with $R^2 = 0.58$. A binomial generalised 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.

Discussion

The significant predicted 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 weak suggestion.

Experiment 3 — relatedness by inclusion

We explored one more way in which premises can be related without matching using set inclusion, as exemplified below:

(11) a. Every guest at this party either owns a car or lives in New York.
Chloe is at the party and lives in the US.
Does it follow that Chloe lives in New York?

b. This party is only for people who own a car and who live in New York.
Chloe is at the party and lives in the US.
Does it follow that Chloe lives in New York?

Method

In this experiment, premise 2 was related to one of the disjuncts of premise one by introducing a superset of one of the disjuncts. In (11), we have "lives in the US," which is a superset of "lives in New York." We translated the previous design as closely as we could to the new set dependencies. Participants were separated in three groups, rating the strength of the inclusion relation, inference-making with disjunction (11a) and inference-making with conjunction (11b).

Participants Participants were 160 individuals in the United States, recruited via Amazon MTurk, 80 for the independent rating and about 40 per condition of the illusory inference. Participants were compensated for this experiment. All 80 participants were kept in the rating task and the inference-making tasks.

Procedure The general structure was as close as possible to the two previous experiments. In the rating task participants responded to questions as in (12) on a 5-point likert scale on a single page. For each element of our materials, participants rated their confidence on either an invalid inference from subset to superset as in (12), or the valid variant from subset to superset (i.e. "the party is for US-residents," "Chloe lives in NY, can she attend?")

(12) This party is for all people who live in New York.
Chloe is at the party and lives in the US.
How confident are you that Chloe is allowed at the party?

In the inference-making tasks the targets had the structure of (11a) for one group and (11b) for the other. The same valid and invalid controls as before were used.

Materials The rating task involved sixteen pairs of sets and their strict subsets. For the inference-making task, from the sixteen pairs eight were randomly drawn to produce target items. The same controls as in the previous experiments were used.

Analysis and results

Rating Participants used most of the upper two thirds of the scale to rate the various set-inclusion inferences. Valid superset to subset inferences received a rating from 0.7 to 0.98, mean 0.87, median 0.90. Invalid subset to superset inferences received ratings from 0.33 to 0.66, mean and median 0.49.

Inference making Figure 3 shows the averaged responses for each item as well as the means across items, and both valid and invalid controls. Binomial tests confirm that in the "and" case, participants' responses to targets were significantly different from chance ($p < 0.001$). In the "or" case, participants' responses to targets were significantly different from their responses to controls: valid $p < 0.001$ and invalid $p < 0.001$. But a binomial test could not rule out the null hypothesis that participants answered randomly to target illusory inferences. The bottom of Figure 3 shows the correlations between the ratings of the crucial invalid subset to superset inferences and the two versions of the illusory inference. The "OR" case yields a significant correlation at the $p < 0.05$ level with a slope of $\beta = 0.74$ ($F[1,14] = 7.05, p = 0.019$), $R^2 = 0.33$. The "AND" case yields a significant correlation at the $p < 0.05$ level with a slope of $\beta = 0.55$ ($F[1,14] = 4.74, p = 0.047$), $R^2 = 0.25$. Correlations with the valid direction of the rating (superset to subset) were not significant.

Discussion

In the "or" condition, participants accepted target fallacies significantly more than invalid controls, which is suggestive

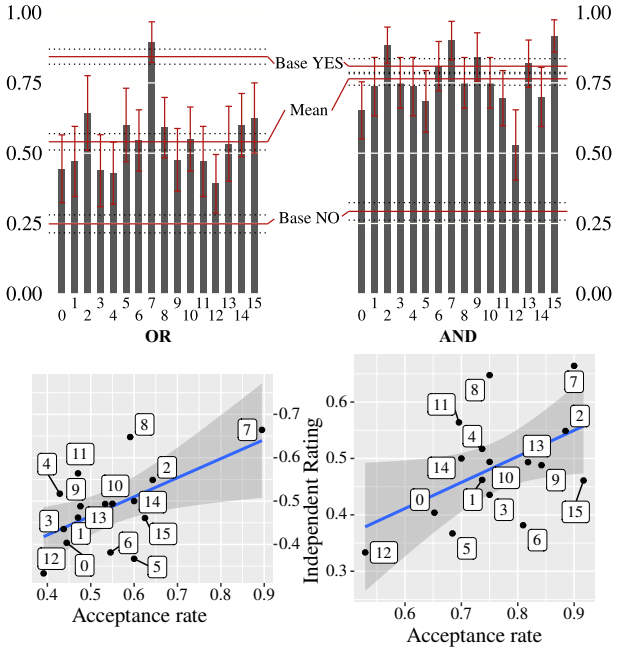


Figure 3: **Top:** vertical bars represent the mean acceptance rate and standard error over the subjects in the "OR" (left) and "AND" condition (right) for each of the 16 stimuli. "Base YES" and "Base NO" show the mean and standard error for respectively the valid controls and the invalid controls. Mean shows the acceptance rate on the fallacy averaged over both stimuli and participants, as well as the standard error. **Bottom:** Correlation between the acceptance rate of the fallacy and the independent rating of set to subset. Correlation for the "OR" and the "AND" against the ratings from the independent experiment.

evidence of the existence of a new illusory inference. Though we could not directly exclude the possibility that participants responded at random to targets, the significant correlation between independent ratings of inclusion relations and responses to targets would be puzzling under the hypothesis that participants responded at chance.

In the "and" version we moved further away from the syntactic structure of the original illusory inference from disjunction: in (11b) we find a conjunction instead. Nevertheless, the conjunction in (11b)-type items still plausibly gives rise to alternative mental models. Premise 1 of (11b) says the party is only for people who own a car and who live in New York. The second a particular individual who may have been at the party, Chloe, is mentioned, it is reasonable that subjects should wonder "which one is she, a car owner, or a New Yorker?" This should be enough to provide alternative mental models, or a question in the sense of the erotetic theory of reasoning, which subjects then try to answer.

However, there is a glaring confound in this conjunctive condition. The first premise of items like (11b) was in fact compatible with two readings. Our intended reading was that the party was for two types of people. But it was possible to interpret that sentence as stating that people had to instantiate *both* properties to be at the party. Under this latter reading, the observed response is no fallacy, it is an absolving interpretation.

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.

In work in preparation (comprehensive handout available at X), we introduce a Bayesian confirmation-theoretic implementation of the erotetic theory of reasoning that extends to the data we report in this article. We propose that, when entertaining competing alternatives as provided by the first premise of illusory inferences from disjunction, reasoners conduct a form of hypothesis-testing. Consider again the example in (1), page 1. Reasoners take the first premise to provide two competing hypotheses, one for each disjunct. Then they seek evidence that can help them decide between the two hypotheses: premise two, on any measure of increase in firmness, constitutes evidence in favor of one hypothesis rather than the other. Checking the extent to which a piece of evidence supports a hypothesis is of necessity a content-sensitive mechanism, so this theory immediately accounts for the novel indirect illusory inferences discussed here. Moreover, this view somewhat surprisingly connects illusory inferences from disjunction with the conjunction fallacy, for which confirmation-theoretic accounts in a similar vein have been proposed (Crupi et al., 2008).

Acknowledgments

The authors would like to thank Benjamin Spector, Emmanuel Chemla, Emile Enguehard, and Philippe Schlenker for very helpful feedback. The research presented here was supported by Agence Nationale de Recherche grants ANR-17-EURE-0017 (FrontCog, Department of Cognitive Studies, ENS) and ANR-18-CE28-0008 (LANG-REASON).

References

Alonso-Ovalle, L. (2006). *Disjunction in alternative semantics*. Phd diss., UMass Amherst.

Crupi, V., Fitelson, B., & Tentori, K. (2008). Probability, confirmation, and the conjunction fallacy. *Thinking & Reasoning*, *14*(2), 182–199.

Cummins, D. D. (1995). Naive theories and causal deduction. *Memory and Cognition*, *23*(5), 646-658.

De Leeuw, J. R. (2015). jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior research methods*, *47*(1), 1–12.

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

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

Koralus, P., & Mascarenhas, S. (2013). The erotetic theory of reasoning: bridges between formal semantics and the psychology of deductive inference. *Philosophical Perspectives*, *27*, 312–365.

Koralus, P., & Mascarenhas, S. (2016). Illusory inferences and the Erotetic Theory of Reasoning. In K. Turner & L. Horn (Eds.), *An atlas of meaning.* Leiden: Brill.

Mascarenhas, S., & Koralus, P. (2017). Illusory inferences with quantifiers. *Thinking and Reasoning*, *23*(1), 33–48. doi: 10.1080/13546783.2016.1167125

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

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