Indirect illusory inferences from disjunction: a new bridge between deductive inference and representativeness

Mathias Sablé-Meyer
UNICOG, CEA, INSERM, Université Paris-Saclay, NeuroSpin, Saclay, France
Collège de France, Paris, France
mathias.sable-meyer@ens-cachan.fr (corresponding author)

Salvador Mascarenhas
Institut Jean-Nicod, Département d'études cognitives
ENS, EHESS, PSL University, Paris France, CNRS
salvador.mascarenhas@ens.fr

Abstract

We provide a new link between deductive and probabilistic reasoning fallacies. Illusory inferences from disjunction are a broad class of deductive fallacies traditionally explained by recourse to a matching procedure that looks for content overlap between premises. In two behavioral experiments, we show that this phenomenon is instead sensitive to real-world causal dependencies and not to exact content overlap. A group of participants rated the strength of the causal dependence between pairs of sentences. This measure proved to be a near perfect predictor of fallacious reasoning by an independent group of participants in illusory inference tasks with the same materials. In light of these results, we argue that all extant accounts of these deductive fallacies require non-trivial adjustments. Crucially, these novel *indirect* illusory inferences from disjunction bear a structural similarity to seemingly unrelated *probabilistic reasoning* problems, in particular the conjunction fallacy from the heuristics and biases literature. This structural connection was entirely obscure in previous work on these deductive problems, due to the theoretical and empirical focus on content overlap. We argue that this structural parallelism provides arguments against the need for rich descriptions and individuating information in the conjunction fallacy, and we outline a unified theory of deductive illusory inferences from disjunction and the conjunction fallacy, in terms of Bayesian confirmation theory.

Preprint of article forthcoming in Review of Philosophy and Psychology (accepted March 2021)

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) P_1: John speaks English and Mary speaks French, or otherwise Bill speaks German.
P_2: John speaks English.
Ccl.: *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. Second, 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.

From an empirical standpoint, this article investigates the matching component of the general account just sketched. We show in our Experiment 1 that examples such as (2) give rise to illusory inferences from disjunction.

(2) P_1: The car slowed down and the guitar was out of tune, or someone was in the attic.
P_2: The brake was depressed.
Ccl.: *Does it follow that the guitar was out of tune?*

Crucially, the second premise does not exactly match any element of the first premise: “The brake was depressed” is certainly *related* to “The car slowed down” but it does not *match* it at a word or content level. In this report (but not in the experiments we conducted), we flag premises that *match exactly* with identical colors, and *related but non-matching* premises with perceptually close colors (red and orange in the example above).

Examples like (2) are central to our argument. As in (1), the conclusion does not follow logically. Indeed, we could be in a situation where “the guitar was out of tune” is false, “someone was in the attic” is true, and “the brake was depressed” is true, making both premises true but the conclusion false. Yet our studies revealed that (i) participants fall prey to this new kind of *indirect* illusory inference from disjunction, and (ii) the extent to which they do so is predicted by a measure of the inferential connection between the two relevant propositions.

Our experimental work establishes two points of interest from a theoretical perspective. The first point casts a shadow on the adequacy of extant theories as accounts of illusory inferences from disjunction. The other shines light on a connection between deductive fallacies like illusory inferences from disjunction and probabilistic fallacies such as the conjunction fallacy or base-rate neglect. This connection had remained hidden for the almost twenty years that the field has known about both kinds of fallacies, but it can now be seen clearly thanks to the novel variants of deductive fallacies that we investigate in this article.

Firstly, it is necessary to revise our best extant accounts of the original illusory inferences with disjunction. These accounts were tailored to examples of the form in (1) on the previous page, where a strict notion of matching was appealing due to its simplicity and to the fact that matching bias of this sort is well documented in the reasoning literature (Evans 1999). By contrast, our novel illusions require a mechanism that is sensitive to *semantic* and *probabilistic* connections between the contents of the propositions involved in the reasoning problem. Specifically, we will argue that both the Revised Mental Model Theory of Khemlani, Byrne, and Johnson-Laird (2018) and the Erotetic Theory of Reasoning of Koralus and Mascarenhas (2013) can in principle account for our new data, though neither currently offers an unproblematic analysis. The New Paradigm in the study of reasoning (Oaksford and Chater 2007) fares no better with our data, or “standard” illusory inferences from disjunction for that matter. We will argue that the most promising strategy to account for our new data is to combine elements from the Erotetic Theory of Reasoning’s semantic approach with a probabilistic element inspired by New Paradigm accounts. While we will not offer in this article a full theory, we will outline one in some detail.

Secondly, our novel indirect illusory inferences as in (2) bear a close structural similarity to problems such as the conjunction fallacy, and related fallacies from the heuristics and biases literature. We will argue for the parallelism in detail in the general discussion. For now, we can summarize it as follows. In the problem in (2) above, a reasoner is presented with two options in the first premise and given some additional information in the second premise. Reasoners seem to notice that the additional information provided by the second premise displays a connection with one of the options from the first premise, the left disjunct, and jump to the conclusion that the left disjunct is in fact true. This is parallel to the conjunction fallacy (Tversky and Kahneman 1983), where reasoners are given two options (bank teller or bank teller active in the feminist movement) and some information (a description). This description is connected to one of the options rather than the other, and reasoners rush to pick that option as being the most probable, in apparent violation of the probability calculus.

There are many differences between the problem in (2) and the conjunction fallacy. Most conspicuously, (A) one of the two options in the conjunction fallacy is included in the other, which is not the case in the two options provided by the first premise of (2); and (B) the additional information used by reasoners to (incorrectly) pick one of the available options in (2) is one brief sentence, while its structural analog in the conjunction fallacy is a full paragraph of background information on an individual, ostensibly building on individuating information and stereotypes. We argue that these sharp differences obscure the structural parallelism, and that acknowledging them is key to understanding the shared reasoning mechanism responsible for both fallacies. Specifically, (A) the inclusion condition between the options in the conjunction fallacy is a key element of how striking a mistake the conjunction fallacy is, but plays no actual role in accounts of the processes that lead to the mistake. Thus, the inclusion relation is not a necessary feature of the mistake, as demonstrated by other representativeness effects such as the lawyers-and-engineers problem (Kahneman and Tversky 1973). Additionally, (B) long descriptions that provide individuating information and rely on stereotypes are not a necessary feature of the conjunction fallacy, or of representativeness effects in general. In conclusion, we will argue that this parallelism between probabilistic problems and our novel indirect illusory inferences from disjunction provides further arguments to unify our field's approaches to deductive reasoning and probabilistic reasoning.

2 Illusory inferences and matching bias

This article focuses on a particular class of illusory inferences from disjunction, discovered by Walsh and Johnson-Laird (2004). In (3) we give a representative example of the fallacies studied in that article, along with its underlying structure.

(3) P_1: $(a \wedge b) \vee (c \wedge d)$ 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.
P_2: a Jane is kneeling by the fire.
Ccl.: b *Does it follow that she is looking at the TV?*

About 85% of subjects judged that the proposed fallacious conclusion followed. Yet it is a fallacy, for it could be that Jane is kneeling by the fire while *not* looking at the TV, and that Mark is at the window peering into the garden. This situation would make the premises true but the conclusion false. Notice that (3) is a fallacy no matter whether the "or" is interpreted exclusively or not.[1] We address a possible absolving interpretation for the first premise in the discussion section, for now we ask any readers already outlining pragmatic explanations to temporarily suspend their disbelief.

The materials used by Walsh and Johnson-Laird (2004) contained four propositions and are unnecessarily complex to address the question we are interested in. Instead we will use the simpler structure in (4), instantiated in (1), with only three propositions.

[1] Formally, an exclusive disjunction in the first premise would amount to $((a \wedge b) \wedge \neg(c \wedge d)) \vee ((c \wedge d) \wedge \neg(a \wedge b))$, which does not validate the inference. The countermodel we present in the main text will do the job here too.

(4)	P_1: $(a \wedge b) \vee c$	(1)	John speaks English and Mary speaks French, or Bill speaks German.
	P_2: a		John speaks English.
	Ccl.: b		*Does it follow that Mary speaks French?*

The structure in (4) 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). Such 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 first clear 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 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 (5) below.

(5) *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.

2.2 Erotetic Theory of Reasoning

Koralus and Mascarenhas (2013) provide a formal deduction system to model naive human reasoning, which takes some inspiration from mental model theory. Their Erotetic Theory of Reasoning incorporates results from linguistic semantics, and recasts the mental models account in terms of a question-answer dynamic. The erotetic theory 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 the erotetic theory 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).

2.3 Shortcomings of exact matching

As they stand, neither theory predicts a fallacy if the second premise fails to exactly match one of the alternatives provided by the first premise, but instead merely displays *a connection* with it. Consider the example in (2) schematized in (6) where independently d and a are connected.

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

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

A conclusion of b in a problem with the structure in (6) cannot be explained by simple matching, since d does not match a. If these kinds of problems are attractive illusions, then our best accounts of *standard* illusions with disjunction as in (1) should be revised. If (2) prompts the same inference-making behavior as (1), then we would ideally want to provide a unified account of both fallacious patterns.

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 experimentally whether participants fell for fallacies based on the structure in (6). We operationalized the link between d and a as causal dependence in order to rely on methodology and materials from experimental work on causal conditionals by Cummins (1995). More specifically, we hypothesized that the perceived strength of the causal dependence between d and a in the schema in (6) above would have a direct effect on the acceptance of the fallacy, and thus set out to measure both independently.

3.1 Method

This experiment required two disjoint studies on two different sets of participants: one to rate the strength of causal dependencies (norming study), and the other to perform an inference-making task on patterns like (6) (Experiment 1).

3.1.1 Participants

We recruited 322 individuals in the United States via Amazon Mechanical Turk to participate in our studies: 242 in the norming study and 80 in the inference task. All subjects were compensated for participating.

Table 1: Breakdown of our participants across the three studies we conducted

	Recruited	Analysed	% Female	Mean age & SD
Norming study	242	238	56.3	32 ± 13.1
Experiment 1	80	64	42.1	34 ± 10.0
Experiment 2	80	70	47.1	34 ± 10.3

Due to a minor wording error regarding the monetary reward in the norming study, we had to halt participant recruitment as soon as we realized the mistake. We corrected the wording and reposted the study. We collected data from 82 participants in the first batch, and 156 out of 160 *new* participants in the second batch: 4 did not report back to Mechanical Turk.

We kept 64 out of 80 participants in the fallacy experiment: 2 did not correctly report back to the Mechanical Turk website and 14 had taken the earlier norming study.

3.1.2 Procedure

Out studies presented themselves as web pages 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 a few demographic questions.

In the norming study 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 below, with repetition of the instructions each time. Two unrelated brief pilot experiments were given in between: a brief Stroop task and a single logic question of a very different nature. They were meant to provide some variety from the repetitive task of judging conditional sentences.

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.

Instructions paid special attention to the deductive nature of the task: "We want to know, once you assume that the sentences you are given are true, whether you think that the proposed conclusion is **guaranteed** to be true" (original emphasis). In the crucial example of deductive invalidity, the conclusion was nevertheless inductively attractive, to bring home the point that participants should zoom in on the right notion of validity.

(7) P_1: At least one student was late to class today.
P_2: When two or more students are late to class, the principal gets worried.
Ccl: The principal was worried.

The instructions said "The correct answer here is **no**. Indeed it's possible that two or more students were late, but all we know for sure is that at least one did. This means that the conclusion is not **guaranteed** to follow."

After the instructions, participants saw *seven indirect illusory inference trials*, structured as in (6) above, presented in random order, and interleaved with *three valid and three invalid controls*. For each trial, participants could answer "yes" or "no" or decide not to answer.

Participants served as their own controls. The role of our control inferences was twofold. Both valid and invalid controls provided a measure of participants' attention and understanding of the task: they were very simple problems with answers not predicted by any theory to be fallacious. Additionally, invalid controls established a crucial baseline for mistakes. Our invalid controls were simple, unrelated to the disjunctive inferences of interest, and not expected to give rise to fallacies. With this design, statistically significant deviations from the baseline for mistakes offered by invalid controls imply fallacious behavior that cannot be explained away as uninformative noise.

Valid controls were instances of modus ponens whose syntactic complexity was comparable to that of the targets and where the correct answer was "yes." 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, and the correct answer was "no." Structurally, P_1: "If a and b, then c," P_2: "not a," "Does it follow that c?"

3.1.3 Materials

We borrowed the causally connected items (a and d in the schema in (6)) from Cummins (1995). The norming task 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.

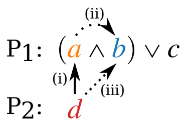

Figure 1: Detailed structure of the premises in (6) and causal dependencies in our target materials. The solid arrow (i) highlights our target dependence while dotted arrows (ii) and (iii) highlight potential confounding dependencies we controlled for in the norming study.

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 kept only those items that showed a moderate or high connection for (i) *d* to *a*, while displaying very weak connections for (ii) *a* to *b* and for (iii) *d* to *b* — as detailed below we had to remove one item from the inference task because of this.

In (8) we give an example of each of the normed connections just discussed. In (9) we give all of our (i) *d* to *a* items.

(8) i. If the trigger was pulled, then the gun fire
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) 0. If fertilizer was put on the plants, then the plants grew quickly.
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 Norming study

Figure 2 shows the ratings of our 8 item sets in the norming study. We report averages across participants together with standard error. Recall that we need for the connections (ii) *a* to *b* and (iii) *d* to *b* to be as low as possible. To assess this requirement we

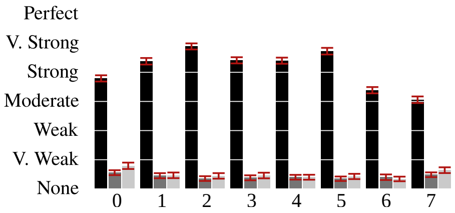

Figure 2: "Strength of the causal connection" rating from the norming study. In black we find block (i) where participants rated $d \Rightarrow a$, in dark gray (ii) $a \Rightarrow b$, and in light gray (iii) $d \Rightarrow b$. Standard error is represented in red.

conducted a one way between-subjects ANOVA to evaluate the effect of the materials on the rating. In block (ii) no significant effect was found at the $p < .05$ level. In block (iii) there was a significant effect at the $p < .05$ level. A post-hoc comparison using Tukey's HSD test indicated that the effect was driven by item 0, which we therefore removed from subsequent experiments. We kept 7 item sets that fulfilled our requirements for the inference task.

3.2.2 Inference study

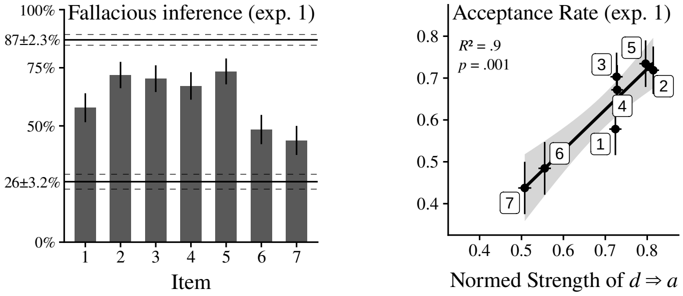

Figure 3: **Left:** Acceptance rate and standard error for each item. The top (respectively bottom) horizontal lines indicate the levels of valid (respectively invalid) controls and the standard errors. **Right:** Correlation between the average reported strength of the crucial dependence d to a from the norming study (x-axis) and the acceptance rate of target fallacies in the inference-making task (y-axis). We plot each individual item. Horizontal and vertical bars indicate the standard error.

Our inference study was a deductive reasoning task, so participants responded to the question "Does the proposed conclusion follow from the premises" with "Yes" or "No." For our first analysis, we considered the rate of acceptance of each item. Valid controls had an average acceptance rate of $87\% \pm 2.3$ while invalid controls had an average acceptance rate of $26\% \pm 3.2$. Figure 3, left panel, shows the acceptance rate per target item with standard error, as well as the average acceptance rate of valid and invalid controls. Figure 3, right panel, shows the correlation between the acceptance rate of the indirect fallacy and the normed strength of the crucial causal connection.

The targets of the inference task are significantly different from chance, valid controls, and invalid controls (Student's t-test, all three comparisons $p < .001$). Acceptance rate in the inference task is significantly predicted by the reported strength of the crucial connection in the norming study ($F[1,5] = 45.0$, $p = .001$). The regression has a slope $\beta = .97$ ($SE = .14$, significant), an intercept of $-.05$ ($SE = .10$, not significant), and a coefficient of determination $R^2 = .9$.

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. This was to check for any effect of attention to the task, as measured by performance in control inferences. Table 2 shows the output of a binomial generalized linear model. We found that the higher the score on controls, the more closely the normed strength of connections predicted acceptance of the target inference (slope approaching 1). This means that the main effect is stronger when subjects perform well on control inferences, showing that the attractiveness of the fallacious conclusion cannot be explained by participants' lack of attention.

Table 2: Best fit of a binomial model that predicts the behavior on the inference task as a function of the normed strength of d to a, participants' score on controls and the interaction between these terms. Normed Strength of d to a and % Error on Controls were centered by removing their mean across subjects. The βs are the coefficients of each term in the binomial regression.

	β	SE β	z-value	p-value
Intercept	0.50	0.10	4.98	< 0.001
Normed d to a	4.14	0.93	4.42	< 0.001
% Error on Controls	-0.37	0.41	-0.91	0.36
Interaction	13.0	3.89	3.35	< 0.001

3.3 Discussion

Our results show that (i) participants find these indirect illusory inferences from disjunction attractive, and (ii) 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.

Studying the interaction between targets and controls shows that some participants are hardly doing the task and give flat answers throughout. Crucially, the greater the accuracy on controls, i.e. the more they are paying attention or the more rational they are, the steeper the slope of the correlation between the normed predictor and the acceptance rate. This indicates that when people are paying attention, they find it easier to resist invalid controls, yet still fall for illusory inferences from disjunction.

The crucial predictor of fallacious behavior is a connection from d to a, which explains over 90% of the observed variance. The connection from d to a relies entirely on world knowledge and cannot be accounted for in terms of matching.

4 Experiment 2 — other forms of indirectness

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

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

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

Experiment 1 showed that the matching component of mental models and erotetic theory of reasoning cannot be the whole story. In that experiment, we leveraged causal dependencies *between premises*. Experiment 2 investigates whether the sensitivity to world-knowledge causal dependencies is restricted to the interaction between premises, or is 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 conclude that a follows by pursuing the causal dependency from d to a?

4.1 Method

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 (6) to (10). We recruited 80 participants, out of which 10 had participated in an earlier related experiment and were removed from the analysis. Subjects were compensated for their participation.

4.2 Analysis and Results

Our inference study was a deductive reasoning task, so participants answered the question "Does the proposed conclusion follow from the premises" with "Yes" or "No." For our first analysis, we considered the rate of acceptance of each item. Valid controls had an average acceptance rate of $90\% \pm 2.1$ while invalid controls had an average acceptance rate of $20\% \pm 2.8$. Figure 4, left panel, shows the acceptance rate per target item with standard error, as well as the average acceptance rate of valid and invalid controls. Figure 4, right panel, shows the correlation between the acceptance rate of the indirect fallacy schematized in (10) and the normed strength of the crucial causal connection measured in the norming study.

The acceptance of targets in the inference task is significantly different from chance, valid controls, and invalid controls (Student's t-test, $p = .0066$ against chance, $p < .001$

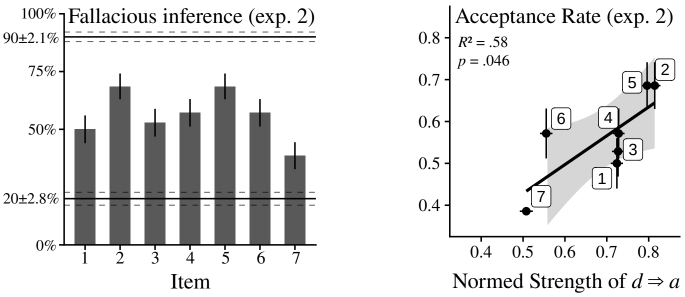

Figure 4: **Left:** Acceptance rate and standard error per item. The top (respectively bottom) horizontal lines indicate the levels of valid and (respectively invalid) controls and the standard errors. **Right:** Correlation between the average reported strength of the crucial dependence d to a from the norming study (x axis) and the acceptance rate of target fallacies in the inference-making task (y axis; exp. 2). We plot each individual item. Horizontal and vertical bars indicate the standard error.

for valid and invalid controls.) Acceptance rate in the inference task is significantly predicted by the perceived strength of the crucial connection d to a in the norming study ($F[1,5] = 6.92$, $p = .046$). The regression has a slope $\beta = .69$, $SE = .26$, and a coefficient of determination $R^2 = .58$.

In a binomial model similar to the one presented Table 2, only the rating of the crucial connection d to a is a significant predictor of the behavior at the $p < .001$ level — with the score on controls and the interaction not having a significant effect.

4.3 Discussion

Experiment 2 confirmed that world-knowledge dependencies aren't only operative in the mechanism that combines the information in the two premises, but also at later steps in the reasoning process. Interestingly, the slope here appears less steep, and less of the variance is explained. This suggests that, while these dependencies are operative throughout, they matter more when one is looking for dependencies between the premises.

5 General discussion

We've established the existence of indirect illusory inferences from disjunction, where the second premise does not properly match any part of the first premise, but instead displays a causal connection to it. 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. Indeed, in matching cases, $d = a$, and thus the central connection from d to a is perfect ($P(a|d) = 1$). The linear model from Experiment 1 then predicts an acceptance rate of $-.05 + 1 \times .97 = .92$ (intercept + 1×slope) for matching cases, in line with the existing literature.[2]

We conclude that a more sophisticated process than exact matching is required, one that is sensitive to contentful connections between models.

5.1 Original mental model theory and the erotetic theory

As discussed in the introduction, Walsh and Johnson-Laird (2004) appeal to a notion of matching to account for classical illusory inferences from disjunction.

It is possible that by "matching" the authors meant a 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 as far as we can tell, an explicit account of such a content-sensitive operation within the original mental model theory does not appear in print. 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.

Similarly, and as described in the introduction, 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.

5.2 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 while 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*. 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.
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 extend an adequate account of *classical*

[2] Our data do not answer the question to what extent non-matching causal connections can approach the acceptance rate found in matching cases. We thank the editor for pointing out this gap. We have shown that our model can successfully predict behavior observed in matching studies, with very high acceptance rates. We further conjecture that materials displaying higher degrees of connectedness than ours should approximate matching cases. The issue is not essential for the theoretical discussion to follow, so we leave it to further research.

illusory inferences from disjunction to the new *indirect* inferences we present in this article.

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.

5.2.1 System 2 processes under low gamma

If the parameter γ is low enough, then *weakly necessary conclusions* will be drawn. A putative conclusion will be a weak necessity of a set of premises just in case (i) every model of the conclusion is included in some model of the premises, and (ii) 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 (4), repeated below as (12).

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

	P1 & P2		
1.	a	b	$\cancel{c}$
2.	a	$\cancel{b}$	c

The conjunction of the premises in classical illusory inferences from disjunction has two mental models: $ab\cancel{c}$ and $a\cancel{b}c$ (a crossed out letter such as $\cancel{x}$ 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 $ab\cancel{c}$), but (ii) there is a model of the premises ($a\cancel{b}c$) 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 by requiring that *every* model of the premises support the conclusion.

However, 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 $a\cancel{b}c$ of the premises, and there is still a model of the premises (now $ab\cancel{c}$) 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) P_1: John speaks English and Mary speaks French, or otherwise Bill speaks German.
P_2: John speaks English.
Ccl. 1: Mary speaks French.
Ccl. 2: Bill speaks German.

Conclusion 1 under (13) is a well-known illusory inference from disjunction, with acceptance rates in the order of 85%. Conclusion 2 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 Conclusion 1. These two reasoning patterns are sharply different, and the System 2 account under low γ just sketched altogether lacks the ability to make this distinction. Interestingly, 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 (13), in fact what is predicted on that theory is a conclusion of *not-c*.

5.2.2 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 (14). Notice in particular the gaps: the first model is silent about *c*, and the second model is silent about *a* and *b*. Nevertheless, since the sentence as a whole is about *a*, *b*, and *c*, the models that constitute its interpretation contain gaps for the propositions they are silent about.

(14)

	P1		
1.	a	b	
2.			c

The second premise of classical illusory inferences has only one simple model: *a*. The next step in the procedure is to conjoin the models of these two premises. This is done by pairwise conjoining each model of premise 1 with each model of premise 2, and collecting all of the *consistent* pairwise conjunctions in a set of alternative mental models.

(15)

	P1				P2
1.	a	b		+	a
2.			c	+	a

The gaps in the models for the first premise are crucial in this process. Following the description of mental model conjunction given by Johnson-Laird and Ragni (2019 Appendix C), the conjunction in the second line of (15) will *not* yield a consistent model. This is because the model of the first premise (*c*) comes with two gaps, one *a*-shaped, the other *b*-shaped. Conjoining a model containing a proposition *p* with a model containing a *p*-shaped gap cannot be done in the theory. Effectively, a *p*-shaped gap in a model behaves as if the model contained *the negation* of *p*, for the purposes of conjunction. Consequently, (15) yields *only one* model, namely *ab*. From here, the observed conclusion *b* follows as a *strong necessity*. The unobserved conclusion *c*, problematic for the System 2 account reviewed in the previous section, does not come out as a prediction of the System 1 account just outlined.

While these predictions are a marked improvement over the predictions under System 2 processes, the account itself is problematic. A central part of getting the right predictions in this System 1 account is the fact that, for the purposes of mental model conjunction, the models of premise 1 work as if they were the models in (16).

(16)

	P1		
1.	a	b	$\not{c}$
2.	$\not{a}$	$\not{b}$	c

This means that the first premise is effectively interpreted in a far stronger way than what (14) suggests. Indeed, the models in (16) for the first premise of the illusory inference from disjunction correspond to what formal pragmatics calls a *strongly exhaustive* interpretation.

These interpretations were first discussed by Spector (2007) in an entirely independent context, as predictions of his theory of scalar implicature, the same mechanism that accounts for exclusive interpretations of simple disjunctions such as $a \vee b$. Extending formal pragmatic methods to the study of deductive fallacies, Mascarenhas (2014) showed that the possibility of interpreting the first premise of the illusory inference from disjunction as in (16) constitutes an absolving interpretation for these fallacies. Under the interpretation in (16), a conclusion of *b* after processing the second premise is no fallacy at all, but a valid inference. Rather than concluding that illusory inferences from disjunction had been entirely misdiagnosed by the mental models literature as fallacies, Mascarenhas (2014) suggested that there are two different paths leading to a conclusion of *b* in illusory inferences from disjunction. The *reasoning path* starts from a straightforward, non-strengthened interpretation of both premises, and delivers a conclusion of *b* as a fallacy via the erotetic theory of reasoning or the original mental-model theory. The *pragmatic path* operates with the strengthened interpretation of the first premise, and derives the same conclusion via entirely valid reasoning.

The original argument in favor of this view was conceptual. There exist inference patterns with disjunction-like operators that give rise to fallacious conclusions highly reminiscent of illusory inferences from disjunction. This has been shown for indefinite quantifiers (Mascarenhas and Koralus 2017) and for the epistemic modal “might” (Mascarenhas and Picat 2019). Both these operators have semantics interestingly connected to disjunction. For example, a sentence with an indefinite quantifier like “A student has arrived” can be seen as a large disjunction “Student 1 arrived, *or* student 2 arrived, *or* . . . ” Crucially, neither indefinite expressions nor epistemic modals can be pragmatically strengthened in a way that would validate their respective illusory inferences. In other words, while illusory inferences *with disjunctions* have two avenues that conspire to render the target conclusion extremely attractive, the *pragmatic path* is not operative in illusory inferences with indefinites or modals. This explained the weaker acceptance rates of the inferences with indefinites and modals, at around 40%.

More recently, Picat (2019) provided strong experimental evidence in favor of the picture suggested by Mascarenhas (2014). Picat has shown that, under cognitive load induced by a concurrent memory task, participants were *less likely* to endorse a conclusion of *b* with illusory inferences from disjunction, while their inference-making behavior was not affected the same way in the case of illusory inferences with indefinites, illusory inferences with modals, or control inferences of an entirely different nature. Importantly, Picat employed a cognitive-load paradigm entirely analogous to one that has been shown to affect the derivation of scalar implicatures, making subjects *less likely*, say, to interpret $a \vee b$ exclusively. These results show conclusively that there is a pragmatically strengthened interpretation of the first premise of illusory inferences from disjunction, but that it is not the full story.

Putting these results together, it is clear that the mental model interpretation of premise 1 in (16) is in fact a pragmatically strengthened interpretation of the premise. That is, the revised mental model theory accounts for illusory inferences from disjunction not as bona fide fallacies, but as the result of pragmatic processes.

At least since Grice (1975), these kinds of pragmatic processes have been argued to follow from communication-specific principles of reasoning. Instead, the revised mental model theory derives at least some scalar implicatures as a matter of entirely domain-general reasoning.[3] This is an intriguing result, but it raises at least three non-trivial questions.

First, what is the view of pragmatics in the revised mental-model theory? At least some cases of scalar implicature arise from the theory as the result of general-purpose reasoning, with no consideration of communicative intents, cooperativeness, or other ingredients of traditional theories of pragmatics. Second, what portion of phenomena traditionally considered to be pragmatic in nature can be handled by mental-model theory? One would most naturally investigate this question by assessing the exact overlap between the strengthening of premises that occurs in mental-model theory at the moment of conjunction and the various proposals in the literature for scalar implicature mechanisms.

Finally, and most importantly for present purposes, the broader picture of illusory inferences becomes quite mysterious. In particular, illusory inferences with indefinites and illusory inferences with epistemics cannot be accounted for in terms of scalar implicatures. This makes it very difficult to see how the revised mental-model theory could give a unified view of illusory inferences from disjunction-like elements. Relatedly, it is hard to see how to understand the experimental results found by Picat (2019).

5.3 Probabilistic theories — the new paradigm

There are very successful ways to account for reasoning under uncertainty with the probability calculus, modeling subjective probabilities (see in particular Oaksford and Chater 2007; Adams 1996; Johnson-Laird, Khemlani, and Goodwin 2015). These theories have been particularly insightful on the broad and important topic of reasoning

[3] There is an alternative view of these interpretive processes from linguistic semantics and pragmatics that considers them to be narrowly grammatical rather than the product of pragmatic reasoning (see for example Chierchia, Fox, and Spector 2012). Since mental model theory is squarely about reasoning, we take it that such a *grammatical* outlook on the extent to which the theory models these kinds of strengthened interpretations was not intended.

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. They have same issue as the revised mental model theory's System 2 account in distinguishing the *b* conclusion from the *c* conclusion. This comes from the fact that the probability calculus includes classical propositional logic, therefore $P((a \wedge b) \vee c, a) = P((a \wedge b) \vee (a \wedge c))$. The latter formula highlights the fundamental symmetry between *b* and *c*, which makes *b* and *c* indistinguishable for this family of theories.

There are at least two ways of operationalizing subjective validity for such theories. The first is *p*-validity, under which a conclusion follows to the extent that it is no less probable than its premises, for every possible probability distribution. Under this view, both conclusions will come out as *not p-valid*: it suffices to define a probability distribution where $P(b), P(c) < \alpha$ but $P((a \wedge b) \vee (a \wedge c)) > \alpha$.

An alternative is to compare not the prior probabilities of premises and conclusions, but the posterior probabilities of the putative conclusions on the premises. Once again this cannot distinguish *b* from *c*. In particular, under the assumption of independent flat 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* will be equally acceptable.

Note that neither of these accounts fare better if they interpret the disjunction as exclusive. For the *p*-validity case, our proposed counterexample still holds and therefore neither conclusion is *p*-valid. In the posterior-driven alternative, $P(b|(a \wedge b) \vee c, a) = \frac{1}{2} = P(c|(a \wedge b) \vee c, a)$ and once again *b* and *c* cannot be distinguished.

5.4 Bayesian confirmation, the erotetic theory of reasoning, and the conjunction fallacy

The indirect illusory inferences from disjunction in this article are surprisingly structurally similar to probabilistic reasoning problems like the conjunction fallacy (Tversky and Kahneman 1983).

(17) "Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations."
Which of these two options is the most probable?

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

Using a number of reasoning problems of which (17) is a particularly well-known example, Tversky and Kahneman (1983) show that reasoners will often assign a higher probability to a conjunction of the form $\varphi \wedge \psi$ than they do to one of its constitutive conjuncts φ, violating the classical probability calculus. To see the structural similarity between (17) and indirect illusory inferences from disjunction, consider that

1. The task in (17) is to pick between one of two options. We can therefore see the choices as inducing a disjunctive premise, telling us that "Linda is a bank teller (b), or she is a bank teller and she is active in the feminist movement ($b \wedge f$)." Reordering the information for ease of exposition, we can say that the task of picking one of those two options induces a disjunctive premise of the form $(f \wedge b) \vee b$.
2. The description of Linda is linked to the proposition f (active in the feminist movement), which occurs in the disjunctive premise just reconstructed. Learning this information about Linda raises the probability that she is active in the feminist movement. Much like the second premise of our indirect illusory inferences "The brake was depressed" raises the probability of "The car stopped," one of the propositions occurring in the disjunctive premise.

In other words, the two options in (17) function structurally like the disjunctive first premise of indirect illusory inferences, setting up the space of possibilities and asking that one be picked. The description of Linda functions as the second premise of indirect illusory inferences, bearing a probability-raising connection to one of the propositions occurring in the disjunction. With this in mind, (18) displays and highlights the structural similarity:

(18)	1. **Indirect illusory inference**	2. **Conjunction fallacy**
	$(a \wedge b) \vee c$	$(f \wedge b) \vee b$
	d, which points to a	d, which points to f
	Ccl.: $a \wedge b$ (whence b)	Ccl.: $f \wedge b$

There are three especially conspicuous differences between the two classes of fallacies, which we argue presently are themselves instructive.[4]

Most importantly, the "disjunctive premise" in the conjunction fallacy contains two disjuncts related by inclusion. $f \wedge b$ entails b, which is not the case with the indirect illusory inference's two disjuncts $a \wedge b$ and c. If our indirect illusory inferences do not display the characteristic set-inclusion via conjunction of the conjunction fallacy, is there any substantive sense in which the two kinds of fallacies are *the same*?

We propose that there are at least two ways in which the conjunction fallacy is a striking and informative datum. First, the conjunction fallacy has the property that the correct answer to it is almost *analytically* available simply by looking at the "disjunctive premise," that is the two possible conclusions. Indeed, a conjunction cannot be more probable than either of its constitutive conjuncts: this means that the correct answer cannot be $b \wedge f$. Assuming that $P(b|f), P(f|b) < 1$, which is certainly true in any conjunction fallacy stimulus we've seen in the literature, one can conclude something even stronger: the individual conjunct *must* be strictly more probable than the conjunction. These conclusions are available without any consideration of the "second premise," the description of Linda. This is indeed a striking property that our indirect illusory inferences do not share: there is no way to decide which disjunct from the first premise of indirect illusory

[4] We are very grateful to an anonymous reviewer, whose detailed comments crucially shaped the discussion that follows.

inferences is more probable, or which disjunct (if any) follows, without considering the additional information in the second premise. This feature of the conjunction fallacy is clearly important, and it has contributed to the enduring appeal of the datum in our field and beyond: when asked directly about it, any naive participant can see that the answer they gave cannot possibly be right, simply by virtue of this inclusion relation between the two options.

But there is another sense in which the conjunction fallacy is interesting and informative, independent from the issue of set inclusion. The conjunction fallacy is an example of reasoning by representativeness, and in this sense the *conjunction* aspect is not a necessary element of the phenomenon. Indeed, Kahneman and Tversky, in the original conjunction fallacy article (1983) and a wealth of other work, saw the conjunction fallacy as a particularly striking special case of a general phenomenon, exemplified also in non-conjunctive problems such as the lawyers-and-engineers paradigm (Kahneman and Tversky 1973), which does not contain options related by inclusion. In the lawyers-and-engineers experiment, Kahneman and Tversky give a description of an individual drawn from a population of lawyers and engineers. The description is much like the kinds of descriptions found in conjunction fallacy experiments, and participants are asked about one of the two atomic categories: how likely is this person to be a lawyer / engineer? Just like in the conjunction fallacy, participants seemingly ignored the probability task, and instead reported a judgment about how typical an example of the category at hand the individual described was. For Kahneman and Tversky, the lawyers-and-engineers and conjunction fallacy tasks are two examples of the exact same reasoning processes. The two problems differ with respect to what is striking about them in senses extraneous to the reasoning process that leads participants to the response they give. In the conjunction fallacy reasoners “should” have been able to see what an egregious mistake their reasoning faculty was pushing them to make, while the same mistake is far more obscure, and perhaps more excusable, in the lawyers-and-engineers task.

Our point in this article is that reasoning by representativeness and reasoning in (indirect) illusory inferences from disjunction is the same process, one we outline at the end of this section. Accordingly, we argue that the conjunction fallacy is parallel to our deductive inferences insofar as the conjunction fallacy is an instance of reasoning by representativeness involving a “disjunctive premise” of the shape $(a \wedge b) \vee c$. The set-inclusion aspect of the conjunction fallacy, which we recognize is of great interest in and of itself, is absent from our indirect illusory inferences, but this does not interfere with the parallelism we are highlighting, which concerns the logical form of the premises, and the mechanisms that lead reasoners to the answers they report, rather than the extent to which it is surprising from a logical perspective that participants should have made the mistake they made.

A second conspicuous difference between the conjunction fallacy and our data concerns the different natures of the two tasks. In the conjunction fallacy participants are asked about probabilities, while in indirect illusory inferences they are asked to perform a properly deductive “is it guaranteed to follow?” task. Much like work in the New Paradigm line, we take it that probabilistic calculations are at the core of reasoning, so that a task ostensibly about logical validity might be best analyzed as involving probabilistic considerations. Thus, it is not in principle incoherent to argue that both the conjunction fallacy (and indeed representativeness reasoning in general) and illusory inferences from disjunction are derived by a faculty of reasoning that is dealing with probabilities, irrespective of whether the dependent measure in the experiments to explain is a direct probabilistic one or a binary decision about validity.

Finally, Kahneman and Tversky considered the presence of individuating information and the reliance on stereotypical reasoning to be essential features of the conjunction fallacy. Our disjunctive illusions rely on world knowledge and thus plausibly on stereotypical reasoning, but by no means do they rely on individuating information. Relatedly, while the original conjunction fallacy design involved an elaborate description of Linda, no doubt in order to best leverage individuating information, our disjunctive illusions display a one-sentence premise performing the same structural job as the description of Linda. This suggests that long descriptions with individuating information are in fact not needed to generate conjunction effects.

Indirect illusory inferences from disjunction and the conjunction fallacy are superficially very different. But if those differences are non-essential from the point of view of developing an account of the two phenomena, as we've argued, then a unified account is in order. One strategy would be to pursue a representativeness account of our indirect illusory inferences. We believe that one is not forthcoming, for our indirect illusory inferences do not rely on individuating information. Representativeness as discussed in the heuristics and biases literature is a three-place relation: an individual i is a representative example of a predicate P to extent d. This kind of notion can be immediately applied to the conjunction fallacy, where we can compare the extents to which two competing predicates are typically exemplified by Linda, the compound predicate "bank teller and feminist," and the simple predicate "bank teller." But *where* is the individual with respect to which to engage in representativeness reasoning, in our indirect illusory inferences from disjunction? Unlike the conjunction fallacy, or indeed any other paradigm in the representativeness literature, our disjunctive premises display three propositions with entirely different subjects and direct objects. Individuating information appears to play no role whatsoever in our problems.

While Kahneman and Tversky considered representativeness and therefore individuating information to be central properties of the conjunction fallacy and related problems, other theorists have provided accounts of the phenomenon of an entirely different nature. In particular, Crupi, Fitelson, and Tentori (2008) argue that reasoners engage in *confirmation-theoretic reasoning* in the conjunction fallacy. Informally, reasoners ask themselves which of the two options (bank-teller or bank-teller-and-feminist) is best confirmed by the available evidence (the description of Linda). There are multiple ways to cash out this kind of reasoning, but a particularly perspicuous one is to check which option's probability is raised the most by learning about the description of Linda. At best, the prior probability of the bank-teller option is unchanged by learning about Linda's engagement with social justice issues. By contrast, the probability of bank-teller-and-feminist goes up, conditional on the same information. It cannot go *above* the posterior probability of bank-teller alone, of course. But it will certainly increase more than its alternative, relative to their respective priors.

We propose that this confirmation-theoretic mechanism is operative throughout human reasoning, certainly in the conjunction fallacy as proposed by Crupi and collaborators, but also almost certainly in representativeness phenomena more generally, and even in deductive reasoning, as per our illusory inferences from disjunction.[5]

It is instructive to see concretely how this confirmation account would work for our data. In each of our materials, the second premise d raises the probability of the disjunct $a \wedge b$, by being causally connected to a, while it is orthogonal to the second disjunct c. Additionally, b and d are unrelated. In probabilistic terms, this means $P(a|d) > P(a)$ and $P(c|d) = P(c)$. Take now a particularly well known measure of confirmation, the Difference measure, where $D(h,e) := P(h|e) - P(h)$; that is, the posterior minus the prior. For the putative c conclusion, $D(c,d) = 0$. But for $a \wedge b$, since b and d are independent by design, it follows from $D(a,d) > 0$ that $D(a \wedge b, d) > 0$ (proof in supplementary materials).

The same holds for all other confirmation measures proposed by Tentori et al. (2007). Confirmation theory can thus account for indirect illusory inferences, if allied to a theory of mental representations that recognizes that disjunctions as those in our first premises put forth two alternatives to decide between. In our view, an *erotetic confirmation* theory of reasoning holds the most promise in this regard. Firstly, because the erotetic theory brings together some of the most important insights of mental model theory and of linguistic semantics, providing an answer to the question of *why* disjunctions bring up alternative possibilities, in terms of its question-answer dynamic. Secondly, because, unlike mental model theory, the erotetic theory is formally amenable to be extended with a probabilistic measure function and therefore Bayesian confirmation tools as proposed by Crupi and collaborators for the conjunction fallacy.

6 Conclusions

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 the new paradigm, 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

[5] Indeed, we find manifestations of this general confirmation mechanism in semantics as well. In particular, recent probabilistic approaches to conditionals propose semantics based entirely on confirmation-theoretic measures (e.g. Crupi and Iacona 2020). Moreover, experimental work by Skovgaard-Olsen, Singmann, and Klauer (2016) shows the influence of confirmation-theoretic considerations on the interpretation of conditionals.

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 are manipulated by attempts to match them with the second premise is an important step toward understanding human reasoning with alternatives.

Importantly, the indirect illusory inferences from disjunction in this article play a useful role connecting the study of failures of deductive reasoning to the study of better known, and perhaps more ecologically valid problems. We argued that the conjunction fallacy can be seen as a special case of our indirect illusory inferences from disjunction. Studying indirect illusory inferences can thus 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 rich descriptions leveraging stereotypical properties. This parallelism creates the need for a unified account of these seemingly disparate fallacies, and can provide arguments of a new nature in favor of extant competing accounts. Representativeness in particular is difficult or impossible to apply to indirect illusory inferences from disjunction, while accounts in terms of confirmation-theoretic reasoning offer compelling and implementable explanations of both data points.

Availability of data and materials

The complete materials, collected data, and analysis code, are available at the following address https://osf.io/tuc8s/.

Acknowledgements

The authors would like to thank Benjamin Spector, Ulrike Hahn, Vincenzo Crupi, Philippe Schlenker, Stanislas Dehaene, Christophe Pallier, Emmanuel Chemla, Lorenzo Ciccione, the LANG-REASON team at Ecole Normale Supérieure's Department of Cognitive Studies, and two anonymous reviewers for helpful feedback.

References

Adams, Ernest W. 1996. *A Primer of Probability Logic*. Center for the Study of Language and Information.

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

Chierchia, Gennaro, Danny Fox, and Benjamin Spector. 2012. “The Grammatical View of Scalar Implicatures and the Relationship Between Semantics and Pragmatics.” In *Semantics: An International Handbook of Natural Language Meaning*, edited by Paul Portner, Claudia Maienborn, and Klaus von Heusinger. Berlin: Mouton de Gruyter.

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.

Crupi, Vincenzo, and Andrea Iacona. 2020. “The Evidential Conditional.” *Erkenntnis*, 1–25.

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.

Grice, Paul. 1975. “Logic and Conversation.” In *Syntax and Semantics: Speech Acts*, edited by P. Cole and J. Morgan. Vol. 3. New York: Academic Press.

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 Marco Ragni. 2019. “Possibilities as the Foundation of Reasoning.” *Cognition* 193.

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

Kahneman, Daniel, and Amos Tversky. 1973. “On the Psychology of Prediction.” *Psychological Review* 80 (4): 237–51.

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*, 549–61. https://doi.org/10.3765/salt.v29i0.4635.

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.

Skovgaard-Olsen, Niels, Henrik Singmann, and Karl Christoph Klauer. 2016. “The Relevance Effect and Conditionals.” *Cognition* 150: 26–36.

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

Tentori, Katya, Vincenzo Crupi, Nicolao Bonini, and Daniel Osherson. 2007. “Comparison of Confirmation Measures.” *Cognition* 103: 107–19.

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. “Co-Reference and Reasoning.” *Memory and Cognition* 32: 96–106.