$(a \wedge b) \vee c$

Inferences with disjunction, interpretation or reasoning?

MSc Thesis

written by
LÉO PICAT

under the supervision of
SALVADOR MASCARENHAS

submitted in partial fulfillment of the requirements for the degree of
MSC IN COGNTIVE SCIENCE

at ÉCOLE NORMALE SUPÉRIEURE

ENS DEC DÉPARTEMENT D'ÉTUDES COGNITIVES

Contents

Acknowledgments **i**

CogMaster's requirements **ii**

Declaration of originality ii

Declaration of contribution ii

Pre-registration iii

1 Introduction **1**

2 Illusory inferences from disjunction 101 **5**

2.1 The first example 5

2.2 Two formal accounts of IIFD 6

2.2.1 The reasoning-based account 7

2.2.2 The interpretation-based account 8

2.2.3 Summary 10

2.3 50 shades of illusory inferences from "disjunction" 10

2.3.1 Class *A* illusory inferences 11

2.3.2 Class *B* illusory inferences 12

2.3.3 Summary 14

2.4 Rationale of the project 14

3 Reasoning load **16**

3.1 Original study 16

3.2 A natural version 17

3.2.1 Methods 17

3.2.2 Results 21

3.2.3 Discussion 22

3.3 An unconnected version 25
3.3.1 Methods 25
3.3.2 Results 26
3.3.3 Discussion 27
3.4 Summary 29

4 Primed reasoning **31**
4.1 Prerequisites 31
4.2 Methods 32
4.2.1 Participants 32
4.2.2 Procedure and stimuli 33
4.2.3 Analyses 35
4.3 Results 36
4.4 Discussion 36

5 Conclusion **39**

Appendix A Might as a generator of alternatives **41**

Acknowledgments

Comme tout le reste sera an anglais, je vais faire ces quelques remerciements dans ma langue maternelle pour pouvoir exprimer comme il se doit ma gratitude aux personnes concernées !

Tout d'abord, un grand merci à mes producteurs et ma maison de disque sans qui rien de tout cela n'aurait pu voir le jour.

Il m'est difficile de remercier Salvador Mascarenhas sans tomber dans l'hyperbole, alors autant y aller franchement ! Au cours de l'année et demie que j'ai eu la chance de passer sous sa supervision, il m'a offert plus d'opportunités d'en apprendre sur le monde académique que je ne puis en compter. Patient, disponible, et engageant, je n'aurais pu rêver d'un meilleur superviseur. Du fond cœur, merci. En espérant que le contenu et le LaTeX de ces pages lui rendent autant honneur qu'il le mérite.

Je tiens à remercier Emmanuel Chemla, Alexandre Cremers, Alejandrina Cristia, et Morgan Sonderegger qui ont su me guider au travers des contrées arides des modèles linéaires à effets mixtes.

Les conseils aiguisées et les idées avisées de Rachel Dudley ont grandement contribué à la troisième partie de ce mémoire. Son aide fut inestimable.

Paul Marty m'a aidé à finement ajuster les paramètres du design expérimental. Il en a toute ma gratitude.

Benjamin Spector, Doreen Georgi, Emmanuel Chemla, Lyn Tieu et Tristan Thommen m'ont donné le goût de la linguistique, merci à eux !

Ces deux années au sein du CogMaster n'auraient pas été les mêmes sans les membres de la fameuse conversation Linguists Assemble et sa séquelle non moins dynamique Linguists Unite.

Enfin, des remerciements sont de mises pour ma famille et mes amis les plus proches amis dont le soutien inconditionnel a été d'un grand support ces dernières années : Maman, Mamie, Anne-Caroline, Tanguy, Nuwan, Abderrahmane, Camille et Juliette.

CogMaster's requirements

Declaration of originality

Illusory inferences from disjunction have received much attention from theoreticians. Their characterization has lead to the development of new formal tools that have proved fruitful in the analysis of well-known reasoning phenomena. On the other hand, the experimental analysis of illusory inferences is a blooming field in which theoretical accounts could be grounded. This work is one of the first trying to show the role of interpretive processes in illusory inferences from disjunction. To our knowledge, the dual-task paradigm and the priming paradigm have never been used on reasoning problems.

Declaration of contribution

Rachel Dudley (RD), Salvador Mascarenhas (SM), Léo Picat (LP) participated in this project. The tasks were distributed as follows:

- definition of the research question — SM
- bibliography — SM
- experimental design — RD, SM, LP
- programming plugins — SM, LP
- programming the experiments — LP
- creation and proofreading of the stimuli — RD, SM, LP
- managing Amazon MechanicalTurk HITs — LP
- data analyses — LP
- data interpretation — SM, LP
- writing the report — LP
- supervision of the writing — SM

Reasoning under cognitive load — Pre-registration

Léo Picat — under the supervision of Salvador Mascarenhas

LINGUAE — Institut Jean Nicod — ENS

This work would be defended at the June session.
The final version of the report would be written in English.
Potential reviewers could be Emmanuel Chemla, Alejandrina Cristia, Christophe Pallier.

Introduction

Background and rationale

Illusory inferences from disjunction

Walsh and Johnson-Laird (2004) first described the illusory inferences from disjunction (IIFD) this proposal focuses on. Canonical cases of it are composed of two premises of the form:

(1) $(a \wedge b) \vee c$
a

About 80% of the participants of Walsh and Johnson-Laird's experiment drew the proposed fallacious conclusion that b is the case. This reasoning is not valid. A situation in which we have a, c and $\neg b$ makes the two premises true and the conclusion false. The fallacy stems from the presence of a disjunction in which one of the disjuncts is a conjunction.

The attractiveness of these fallacies calls for a formal explanation of their emergence.

Reasoning-based account of the illusory inference from disjunction

Mental models theories are the only reasoning-based accounts of the illusory inferences from disjunction. Among them, the erotetic theory of reasoning (Koralus and Mascarenhas, 2013) proposes that reasoning is a way to answer questions using each premises in a systematic manner to maximize their communicative utility.

The gist of it is the following: $(a \wedge b) \vee c$ asks if we are in a $a \wedge b$ or in a c situation. The second premise a overlaps with $a \wedge b$ and drives the conclusion that we are also in a b situation.

Interpretation-based account of the illusory inference from disjunction

On the interpretation-based account, illusory inferences do not follow from incorrect inference patterns but are the result of complex pragmatic processes. What looks like a failure of reasoning is instead the result of entirely justifiable interpretative processes.

For the case at hand, scalar implicatures are the relevant notion, starting from the accounts offered by Sauerland (2004) and Spector (2007). In a nutshell, the presence of the disjunction triggers a scalar implicature that will strengthen the original $(a \wedge b) \vee c$ into $(a \wedge b \wedge \neg c) \vee (c \wedge \neg a \wedge \neg b)$. Taking this strengthened meaning as the first premise, it is a *valid inference* to conclude b from the second premise a.

There are multiple versions of IIFD, presented in section* 1. The two lines of explanation do not account both for every versions: some of them have an interpretation-based account, some do not. This should make the former, and not the latter, sensible to pragmatic manipulations.

Project

The interpretation-based and the reasoning-based approach both offer a formal account of a failure of reasoning. This raises the challenge to understand to which extent each theory can explain the phenomenon. The complexity of the interpretation-based approach should make it sensitive to distraction such as cognitive load or priming whereas the reasoning-based approach proposes a more automatic process.

Key research question

This project is integrated into the bigger picture of understanding the line between reasoning and interpretation processes.

The research question can be articulated as follow: What are the effect of pragmatic processes manipulations, namely manipulating the rate of scalar implicatures, on different versions of the illusory inference from disjunction?

General hypotheses

We expect more logical reasoning, i.e. fewer fallacious conclusions when pragmatic processes are reduced, for IIFD for which a interpretation-based account have been proposed.

Methods

Different illusory inferences from disjunction

Illusory inferences from disjunction come in different flavors. They all rely on a disjunctive-like element. They belong to one of two types depending on whether or not an interpretation-based account have been proposed for them.

IIFD with an interpretation-based account

The canonical example is of the form $(a \wedge b) \vee c$.

(2) Either Jane is kneeling by the fire and she is looking at the TV or otherwise Mark is standing at the window.
Jane is kneeling by the fire.
Does it follow that she is looking at the TV?

One can paraphrase the pragmatic enrichment by adding *only*. The first premise is then to be understood as *only* $a \wedge b$ *or else only* c. Given the second premise b, it follows that a.

Another version replaces the conjunction by a universal quantifier. Universal quantifers can be seen as potentially infinitary conjunctions over the domain of individuals. These variants are of

the form $\forall x.P(x) \vee \forall x.Q(x)$

(3) Mary has met every king or every queen from Europe.
Mary has met the king of Belgium.
Does it follow that she has met the king of Spain?

After pragmatic enrichment, the first premise is to be interpreted as *Mary has met only every king of Europe or only every queen of Europe*. Here as well the conclusion follows from the premises.

This type of IIFD should be sensible to pragmatic manipulations.

IIFD without an interpretation-base account

The presence of a disjunction is not always necessary to trigger an illusory inference from disjunction.

A first example is of the form $\exists x.P(x)$

(4) Some pilot writes poems.
John is a pilot.
Does it follow that John writes poem?

The yes answer is also attractive in (4) (Mascarenhas and Koralus, 2017). Here, the disjunction is hidden within the existential quantifier, which can be interpreted as an infinitary disjunction of every pilot. Here pragmatic enrichment would result in *Some but not all pilots write poems*. The conclusion does not follow from the premises.

In the run-up stage in the Fall, we discovered a new kind of illusory inferences from disjunction where the disjunction is replaced by a modal. This results in the following form $\Diamond(P(x) \wedge Q(x))$

(5) Miranda might play the piano and be afraid of spiders.
Miranda plays the piano.
Does it follow that Miranda is afraid of spiders?

Here as well, no pragmatic enrichment can account for the fallacious conclusion.

This type of IIFD should not be sensible to pragmatic manipulations.

In sum, there are two types of IIFD:

- the canonical and the universal ones, later referred to as IB, which have an interpretation-based account;
- the existential and the modal ones, later referred to as not-IB, which lack an interpretation-based account.

Indirect evidence supports this division into two classes: the first class for which an interpretation-based account have been proposed have higher acceptation rate (80%) than the second class (50%).

We will run two different experiments, each of them will use a different pragmatic manipulation: the first one will be based on cognitive load, the second one on priming.

Cognitive load

De Neys and Schaeken (2007) successfully used a dual-task paradigm to reduce the rate of scalar implicatures processed by participants. We want to translate this methodology to an inference task to see if the effect reduces pragmatic interpretations of the crucial first premise, thereby blocking interpretation-based routes to the fallacious conclusion. This study was also pre-registered on Open Science Framework under embargo.

Participants

We will recruit 170 American participants on Amazon Mechanical Turk. For this decision we consulted with Paul Marty, a researcher at ZAS Berlin who has ongoing research on the dual task paradigm as a blocker of scalar implicatures. We also know that we are likely to exclude around 30% of our participants given the pilot we ran.

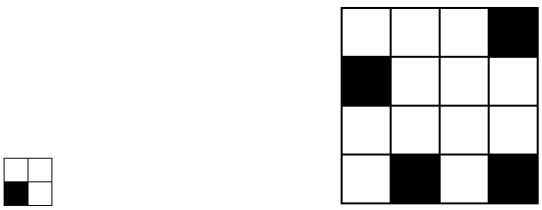
We will exclude participants with a background consisting in more than one graduate course in natural language semantics or pragmatics. We will exclude participants who fail to answer correctly to more than 25% of the controls. We will exclude participants who did not respond to more than 50% of the questions. We will exclude participants who reported using lots of notes or diagrams during the task.

Procedure and stimuli

We will use a dual-task paradigm in a 2×2 within-subject design: we will manipulate the difficulty of the task and the type of IIFD presented. The first task will be to remember patterns of black squares on a $n \times n$ grid. The second task will be to decide if a conclusion follows from a given set of premises.

A fixation cross will be displayed during 1s. The participants will be shown a pattern of black squares during 850ms. They will then evaluate an inference. Finally, they will be invited to reproduce the previous pattern on a blank grid.

In the easy condition, the grid dimensions will be 2×2 and the pattern will consist of a single black square, randomly generated. In the hard condition, the grid dimensions will be 4×4 and the pattern will consist in: either 4 unconnected black squares such that there is never a black square in the Moore neighborhood of another black square, either 2 unconnected black squares and 2 connected black squares. The Moore neighborhood of a given square is composed of the 8 squares around it. The exact patterns are randomly generated.



(a) Easy condition (b) Hard condition

Figure 1 – The memory-load task

Following extensive piloting and the recommendations of our Berlin collaborators, participants will be first presented the hard condition.

Examples of controls and targets inferences are given in table 1. Participants will see a total of 24 items per condition (8 controls and 16 targets), in a random order. Items are gathered into two groups of 24 items. The group participants will see first is randomly chosen. We selected the items that displayed the lowest connection between the second premise and the conclusion. We based our decision on pilot data in which we asked participants to evaluate specifically the strength of the connection between sentences. Two native English speakers proofread the final list of items.

Target IB — Canonical	Sue speaks English and Peter speaks Japanese, or else Jane speaks Spanish. Sue speaks English. *Does it follow that Peter speaks Japanese?*
Target IB — Universal	Mary has met every king or every queen from Europe. Mary has met the king of Belgium. *Does it follow that Mary has met every king of Europe?*
Target not-IB — Existential	Some pilots writes poems. John is a pilot. *Does it follow that John writes poem?*
Target not-IB — Modal	Miranda might play the piano and be afraid of spiders. Miranda plays the piano. *Does it follow that Miranda is afraid of spiders?*
Yes-filler — MP	If everyday is rainy, then every kid will stay home. Everyday is rainy. *Does it follow that every kid will stay home?*
Yes-filler — DS	Bob is chatting with Emma, or else Harry is drinking coffee. Harry is not drinking coffee. *Does it follow that Bob is chatting with Emma?*
No-filler — Fake MP	If Carl dies his hair, Lily will be delighted. Franklyn has a new car. *Does it follow that Lily will be delighted?*
No-filler — Fake DS	Jack bought every Harry Potter books or every Star Wars movies. Jack bought every Harry Potter books. *Does it follow that Jack bought every Star Wars movies?*

Table 1 – Examples of stimuli for the dual-task experiment (MP = *modus ponens*, DS = disjunctive syllogism)

Measures

Regarding the first task, we will measure the mean number of perfectly reproduced grid per condition.

Regarding the second task, we will measure:

- the answers to the controls (correct or incorrect);
- the answers to target inferences (accepted and unaccepted), for each type of them.

Predictions

We expect participants to accept the fallacious conclusion less often in the hard condition compared to the easy condition, only for IIFD for which an interpretation-based account have been proposed. We do not expect any change in the rate of acceptance of IIFD for which no interpretation-based account have been proposed.

Analyses

We plan to analyze our data using binomial linear mixed effects model:

- the dependent variable will be the answer to target inferences (accepted or not);
- the fixed effects will be the condition (hard or easy), the type of IIFD (interpretation-based (IB) or not (not-IB));
- the by-subject random effects will include a random intercept, a random-slope for condition and a random-slope for type of IIFD;
- the by-item random effects will include a random-intercept and a random-slope for condition.

We'll first use Helmert coding.

If the model fails to converge or exhibits a singular fit, we'll use a different coding strategy (sum coding). If this does not solve our issues, we'll try to remove correlations between random effect or remove random effects until the model behaves normally. We'll assess the effect of these simplifications by comparing the final model with the full model. We'll use post-hoc tests to test our predictions with a multivariate t distribution to correct for multiple comparisons.

Interpretation

Using post-hoc tests, we expect a significant difference between the easy and the hard conclusion for IB but not for not-IB. Based on previous work, we also expect the difference between IB and not-IB to be significant at least in the weak condition.

Priming

Bott and Chemla (2016) successfully used a priming paradigm to manipulate the rate of pragmatic enrichment processed by participants. They show that it is possible to prime subjects to enrich sentences containing a given scalar item with sentences containing another scalar item.

IIFD interpretation-based account relies on the enrichment of a sentences containing a disjunction. This type of scalar item was not investigated in the work cited above. We first describe the final experiment we would like to run and then the different pilots that need to be run before that.

Participants

We will recruit 200 American participants on Amazon MechanicalTurk.

We will exclude participants with a background consisting in more than one graduate course in natural language semantics or pragmatics. We will exclude participants who fail to answer correctly to more than 25% of the controls. We will exclude participants who did not respond to more than 50% of the questions.

Procedure and stimuli

We will use a priming paradigm in a 2×2 between-subject design: we will manipulate the type of priming and the type of IIFD presented.

The experiment will consist in a succession of triplet of trials: two prime trials and a probe trials.

Priming trials In the priming trials, we will prime participants to interpret sentences containing a scalar term pragmatically (strong reading) or logically (weak reading). Participants will perform a truth-value judgment task to decide if a picture is a good match for a sentence. They will be presented a picture composed of colored UNICODE symbols and a sentence. They will be informed that the picture has been chosen by another person to match the sentence. They will then assess if this is a good match for the sentence. The priming will be achieved through crucial items (a sentence containing a scalar term and a picture corresponding to the weak reading of the sentence).

Participants will be given feedback on their answer. We will use different feedback to prime participants to interpret sentences in a weak or strong manner (weak or strong condition). The feedback sentences have been proof-read by two native English speakers.

The scalar item we will use is the disjunction. The crucial sentences will be of the form *There is a SYMBOL-1 or a SYMBOL-2*. The picture will consist of the two symbols mentioned. In the weak condition, the feedback will point that it was a good match. In the strong condition, it will point that it was not.

A complete description of the items and the feedback sentences is given in table 2 and 3.

Crucial item	★ ♣	There is a star or a club
Yes-filler	★ ♥	There is a star or a club
Yes-filler	★ ♣	There is a star and a club
No-filler	★ ♥	There is a star and a club
No-filler	♦ ♥	There is a star and a club
No-filler	♦ ♥	There is a star or a club

Table 2 – Stimuli for the priming phase — Disjunction version

Crucial item	Answer given is "Yes"	Answer given is "No"
Weak condition	Yes, that's a good match.	Wait. That's a good match.
Strong condition	Wait, that's a bad match	Yes that's a bad match.

Table 3 – Feedback for the priming phase — Disjunction version

Probe trials In the probe trials, participants will decide if a conclusion follows from a given set of premises. The material used will be the same as in the dual-task paradigm.

Participants will be assigned to one of the two following conditions:

- "strong" condition, in which the priming phase will prime strong readings;
- "weak" condition, in which the priming phase will prime weak readings;

Measures

We will measure:

- the answers to the controls (correct or incorrect);
- the answers to target inferences (accepted and unaccepted), for each type of them.

Predictions

We expect participants to accept the fallacious conclusion more often in the strong condition compared to the weak condition only for IIFD for which a interpretation-base account have been proposed. We do not expect any chance in the rate of acceptance of IIFD for which no interpretation-based account have been proposed.

If our experiment is powerful enough, we have two more predictions. We expect participants to accept the fallacious conclusion more often in the strong condition compared to the control condition. We expect participants to accept the fallacious conclusion less often in the weak condition compared to the control condition.

Analyses

We plan to analyze our data using binomial linear mixed effects model:

- the dependent variable will be the answer to target inferences (accepted or not);
- the fixed effects will be the condition (strong or weak priming), the type of IIFD (interpretation-based (IB) or not (not-IB));
- the by-subject random effects will include a random intercept, a random-slope for condition and a random-slope for type of IIFD;
- the by-item random effects will include a random-intercept and a random-slope for condition.

We'll first use Helmert coding.

If the model fails to converge or exhibits a singular fit, we'll use a different coding strategy (sum coding). If this does not solve our issues, we'll try to remove correlations between random effect or remove random effects until the model behaves normally. We'll assess the effect of these simplifications by comparing the final model with the full model. We'll use post-hoc tests to test our predictions with a multivariate t distribution to correct for multiple comparisons.

Interpretation

Using post-hoc tests, we expect a significant difference between the easy and the hard conclusion for IB but not for not-IB. Based on previous work, we also expect the difference between IB and not-IB to be significant at least in the weak condition.

Follow-up

If we have time, we will run another version of this experiment. We will use a quantifier in priming trials. The crucial sentences in the priming trials will be of the form *Some of the SYMBOL are COLOR*. The picture accompanying the sentence will be the same symbols displayed twelve times in the given color. In the weak condition, the feedback will point that it was an appropriate description. In the strong condition, it will point that it was not. Table 4 provides a description of the items we will use. The feedback will be idle to the previous version of the experiment.

Crucial item	★★★	Some of the stars are red
Yes-filler	★★★	All of the stars are red
No-filler	★★★	All of the stars are red
No-filler	★★★	All of the stars are red
No-filler	★★★	Some of the stars are red
No-filler	★★★	Some of the stars are red

Table 4 – Stimuli for the priming phase — Quantifier version

Piloting

As explained above, we need to ensure the reliability of the priming paradigm regarding disjunction. We plan a succession of pilot studies to determine if the paradigm can be used regarding IIFD.

Step 1 ($\vee \rightarrow$ P1) is to determine if we can prime enrichment of sentences containing a disjunction. We will use a disjunction as the scalar item in the priming trials and sentences of the form $(a \wedge b) \vee c$ (lately referred to as P1) in the probing trials. If this does not work, there is no reason to expect that the priming paradigm can be used for our purpose. If step 1 is a success, we will address step 2 and 3.

Step 2 ($\vee \rightarrow$ IIFD) corresponds to what we describe above using disjunction in the priming trials.

Step 3 is divided in two parts, 3a and 3b. In part 3a ($\exists \rightarrow$ P1), we will determine if the enrichment of P1 sentences can result from priming trials using quantifier. If part 3a is a failure, we will look if quantifier priming can prime disjunction enrichment (step 4 ($\exists \rightarrow \vee$)). If this is not the case, we will conclude that the priming paradigm is not suitable for our purpose. If part 3a is a success, we will turn to part 3b.

Part 3b ($\exists \rightarrow$ IIFD) corresponds to what we describe above using quantifiers in the priming phase.

The different steps are presented in figure 2.

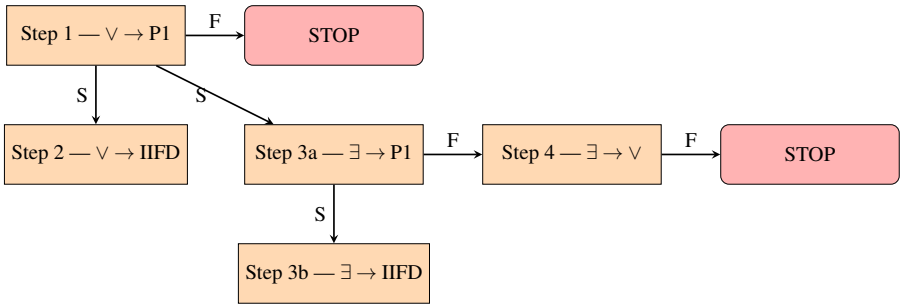


Figure 2 – Pilot plan for the priming experiment. S stands for the success of a given step and F for its failure

Step 1 The interpretation-based account of IIFD relies on the strengthening of P1 sentences. We will first check if the priming paradigm can lead to strengthen such sentences in a truth-value judgment task.

We will use sentences based on a disjunction in the priming phase. Participants will perform a

similar task in the testing phase: they will determine if a sentence is an appropriate description for a picture. The target sentences will be of the type P1 in the testing phase and the picture will correspond to a weak reading of the sentences. The exact stimuli we will use still need to be determined.

Participants will be assigned to one of the three conditions described above.

We will measure:

- the answers to the controls (correct or incorrect);
- the answers to P1 items (appropriate description or not).

We expect participants to accept the description more often in the weak condition compared to the strong condition.

We will analyze the data as described above, except the model will not include fixed and random effects associated to the type of IIFD (this factor being not present here).

If these predictions are borne out, we will consider step 1 to be a success.

Step 3a If step 1 is successful, we will check if the priming paradigm is efficient across the scales we are interested in, i.e. we will check if priming participants with sentences containing a quantifier can affect the interpretation of sentences of the type P1.

The design and material will be the same as described before.

The measure and predictions are the same as before.

If these predictions are borne out, we will consider step 3a to be a success.

Expected contributions

Rachel Dudley (RD), Salvador Mascarenhas (SM), Léo Picat (LP)

- definition of the research question — SM
- bibliography — SM
- experimental design — SM, LP
- programming plugins — SM, LP
- programming the experiment — LP
- creation and proofreading of the stimuli — RD, SM, LP
- recruiting of the participants — LP
- data analyses — LP
- data interpretation — SM, LP
- writing the report — LP
- supervision of the writing — SM

Chapter 1

Introduction

What makes us humans different from the rest of the living realm? Walking on two feet? Kangaroos do as well. Having an opposable thumb? So do gorillas. Not having fur on our skin? Some ugly rats do not either. A better way to pinpoint what is unique about us is not to focus on what we look like but on how we behave. Indeed humans engage in extremely diverse behavior with one another: the complexity of our interactions, describing and understanding is a major focus of several disciplines, some of which have emerged recently to focus entirely on these questions (social psychology, sociobiology for instance).

Reasoning is a key feature that underlies and makes these interactions possible. Even though what might pop into one's mind when reasoning is at issue are complex mathematical proofs or long philosophical dissertations, reasoning is actually involved in daily life. Reasoning can be defined as the faculty that allows us to draw inferences from previous information. Following this view, it becomes clear that reasoning is ubiquitous in our lives: understanding that your spouse is upset from the frown of his or her brow, deriving that a mosquito bit you because of the itchy welt on your arm, or guessing it rained from the puddles on the road.

Let's put our reason into action with the following problem:

(1) Mary has met every king or every queen of Europe.
Mary has met the king of Belgium.
Did Mary meet the king of Spain?

Innocent at first glance, problems like this three-line example will be the main focus of the 1732 lines of this thesis. What is interesting about it? The intuitive answer to the problem seems to be that conclusion follows, end of story, let's talk about something more interesting. However, this is not a sound conclusion: a situation in which Mary met every queen of Europe, the king of Belgium and no one else, makes the two premises true and the conclusion false. The truth of the premises does not guarantee the truth of the conclusion. Yet, the wrong answer is attractive. This grants it the title of reasoning failure.

Reasoning failures challenge a long-lasting view of human reasoning as a means to take the best possible decision in any given situation. They exhibit clear limits to our reasoning capacities and have been a major focus of psychology of reasoning over the past centuries. Psychologists discovered many of them and performed many experimental studies to characterize the different fallacies and determine what triggers them. Conceptually, reasoning can fail in two distinct ways: accepting a false conclusion and refusing to accept a valid conclusion. The former are called compelling fallacies and the later repugnant validities. Compelling fallacies have historically received a major focus and this thesis will give them even more. In what follows, *reasoning failures* and *reasoning fallacies* will mainly refer to compelling fallacies. We give here a small list of deductive problems to give the reader a glimpse of the variety of the field.

Affirming the consequent (Rips, 1994)

(2) If you can make a full audience laugh, you are funny.
You are funny.
Therefore you can make a full audience laugh.

You could be funny because you make good jokes, but only in small groups.

Denying the antecedent (Rips, 1994)

(3) If you travel in business class, you are rich.
You don't travel in business class.
Therefore you are not rich.

You could be a billionaire and yet not travel in business class because you have your own plane.

Wason selection task (Wason, 1968)

(4) The following cards are in front of you: E, K, 4, 7. Each card has a letter on one side and a number on the other side. The rule is that if the card has a vowel on one side, it has an even number on the other side. Which cards must you turn over to check if the rule is verified?

Half of the people respond E and 4, yet the correct answers are E and 7. You have to turn E to make sure there is an even number on the other side, but the rule says nothing about what letter should be on the other side of a card with an even number. The rule also says nothing about what number should be behind K. However, if there were an E on the other side of 7, the rule would be falsified.

Illusory inferences from disjunction (Mascarenhas, 2014)

(5) Mary has met every king or every queen of Europe.
Mary has met the king of Belgium.
Therefore Mary has met the king of Spain.

The presence of the word *or* is suspected to play a big role in the attractiveness of the fallacious conclusion. Thus, this type of reasoning failures is called illusory inferences from disjunction.

One lesson to draw about the field of psychology of reasoning based on these examples is its heavy reliance on linguistic stimuli.[1] To develop and constrain their theories, psychologists gather data. These data are largely produced through behavioral experiments. During these, participants are presented reasoning problems framed with natural language, most of the time English. Surprisingly, this field has not often sought the expertise of linguists to shed new light on their materials.

Semanticists would have been particularly useful. Semantics is the field of linguistics dedicated to understanding how meaning emerges from sentences. Over the years, semanticists have developed sophisticated theoretical frameworks to capture what a sentence means. To do so, they draw inspiration from formal logic. Thus, logical entailments and sound reasoning have been a major focus for semanticists interested in reasoning. Surprisingly again, this field has ignored most of the points of interest that arise in reasoning failures.

On the other hand, linguistics in a broad sense and psychology have successfully collaborated in the past to uncover new interpretations of behavior previously classified as fallacious. Piaget's groundbreaking studies on infants' reasoning exemplify this: it had long been thought that young children drastically lacked object permanence (when an object is not in their sight, they assume it does not exist). Insights from pragmatics brought by Topál et al. (2008) showed that this behavior is partly conveyed by infants' desire to be a cooperative partner: remove human interaction from the task and children get more logical. This example shows how the two disciplines can benefit from each other.

These issues have been brought to the attention of the scientific community by a recent line of research (Mascarenhas, 2014), which proposed to address them by exploring new points of contact between linguistics and the psychology of reasoning. This approach brings the tools offered by semantics into the analysis of reasoning failures. The work presented expands work within this research program on illusory inferences from disjunction.

Because these problems are presented with language, a reasoner giving an answer successively performs two distinct tasks. First, she needs to decode the linguistic material

[1] Of course, some work has made use of non-linguistic stimuli, especially in developmental studies. For instance, Mody and Carey (2016) and Cesana-Arlotti et al. (2018) have studied the emergence of disjunctive syllogism in infants. However, reasoning problems with high structural complexity such as would be addressed in this thesis have not been studied. Ongoing work in our group is developing new experimental paradigms to explore illusory inferences from disjunction in children. Results should come in soon.

and form mental representations of the premises. Second, she manipulates these mental representations to draw new conclusions. Several accounts have been proposed for illusory inferences from disjunction. They vary as to whether they identify the source of the mistake in the first step, the interpretation of the premises, or in the second step, the reasoning made with the premises. Theoretical work has explored these two steps in detail (Koralus and Mascarenhas, 2013; Mascarenhas, 2014, 2013; Mascarenhas and Koralus, 2017) shedding light on the line between interpretive processes and general-purpose reasoning. On the other hand, experimental strategies have been less successful in tearing apart the role of these two different potential routes to the fallacious conclusion.

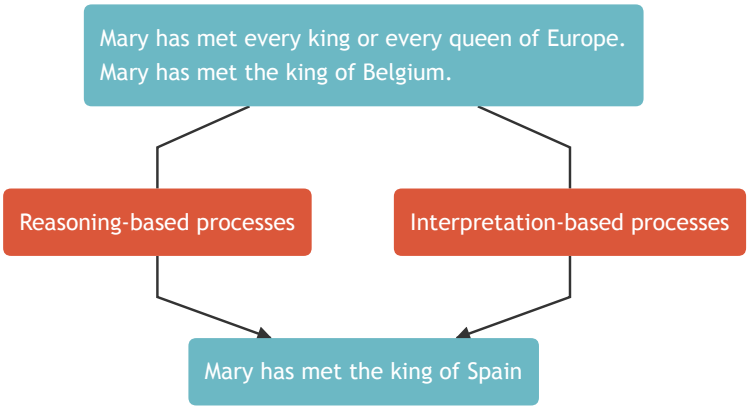


Figure 1.1 – Reasoning and interpretation conspire in driving fallacious conclusions, both making a contribution to an heterogeneous phenomenon.

In this work, I propose experimental paradigms to put forth the involvement of interpretation-based processes in the generation of illusory inferences from disjunction. I use cognitive load as a specific intervention to impair interpretation-based processes. This results in a decrease in fallacious conclusions in response to illusory inferences from disjunction. This is the first direct experimental proof of an interpretation-based account of these fallacies.

The first section provides an extensive introduction to illusory inferences from disjunction. After presenting the discovery of the phenomenon, I will review two theoretical accounts of the fallacy. Then I will provide a list of illusory inferences from disjunction-like elements. The second section reports the implementation of a dual-task paradigm using cognitive load as a way of uncovering the relative contribution of the processes I would have presented in the first section. Finally, the third section presents an ambitious priming paradigm and its potential application to the study of illusory inferences from disjunction.

Chapter 2

Illusory inferences from disjunction 101

The glimpse of illusory inferences from disjunction we had in the introduction does not do justice to their underlying complexity and the various forms they can take. This section paints a more complete picture of the phenomenon. I will first describe the canonical case of illusory inferences as they were first discovered. Then I will present two accounts of the fallacy. Finally I will provide a list of variants of illusory inferences. (Hereafter I will sometimes refer to illusory inferences from disjunction as IIFD.)

2.1 The first example

Walsh and Johnson-Laird (2004) first discovered these illusory inferences from disjunction. In their study, they proposed the following inference to their participants.

(6) 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?

80% of their participants drew the proposed conclusion. It is perspicuous to consider the schematic structure of the inference above:

(7) $(a \wedge b) \vee (c \wedge d)$
a
b?

Concluding that b is unsound reasoning: a situation in which c, d and a are true and b is false would make both premises true but the conclusion false. Yet, in addition to being attractive, the fallacy is robust to several variants: whether a, b, c and d involve different subjects or not, are compatible with each other or not makes no difference: participants

still exhibit fallacious behavior by concluding that b follows.[1]

Walsh and Johnson-Laird proposed an account of the fallacy relying on mental models (Johnson-Laird, 1983; Walsh and Johnson-Laird, 2004). Here I sketch a derivation of the fallacious conclusion using a modified version of mental models integrating some elements of the erotetic theory of reasoning (detailed below). The first premise induces two mental models, one corresponding to each disjunct: one in which a and b are the case and one in which c and d are the case. The second premise is related to the first model (a and b) and not the second (c and d). The second model drops from attention and the reasoner is left with a and b, whence b follows. This is sufficient to trigger the fallacious conclusion.[2]

This account has the advantage of being intuitive. It allows an easy understanding of what cognitive processes are hypothesized by the theory of mental models. Furthermore, it is part of a highly successful theoretical framework in the study of human reasoning (Johnson-Laird, 1983) Nevertheless, mental models theory as presented and used in Walsh and Johnson-Laird (2004) lacks formal rigor as noted by Hodges (1993). In this regard, mental models pale in comparison to semantic theories.

2.2 Two fully-formal accounts of illusory inferences from disjunction

I will first present an alternative account still relying on mental models but building on a set of fully explicit operations. Then I will focus on a competing account that uses tools borrowed from semantics and pragmatics. I call these the reasoning-based account and the interpretation-based account, respectively, in terms I will define later. The way each of these accounts derives the fallacious conclusion will be exemplified on a simplified version of the illusory inferences presented previously.[3]

(8) P_1: $(a \wedge b) \vee c$
P_2: a
C: b?

[1] The content can make a significant difference between variants. For example, a version in which a, b, c and d feature the same agent yields more mistakes than a version in which they each feature a different agent. However, the content does not make a difference in the sense that each variant drives participants to draw fallacious conclusions. Furthermore, the difference between the same-agents version and the different -agents version is small. The effect is not an artifact of the number of agents.

[2] This fallacy is not due to the fact that the conclusion is explicitly proposed to participants. Mascarenhas and Koralus (2016) replaced the final question with a free form introduced by *What, if anything, follows?* They obtained similar responses in favor of b. More interestingly, participants were as confident in their answer irrespective of the way they were prompted to give it (either an explicit proposition or a free form).

[3] The derivations for the version used in Walsh and Johnson-Laird (2004) will be left to the reader if she feels brave enough.

2.2.1 The reasoning-based account

The reasoning-based account identifies the source of the fallacious behavior in the way premises are combined with each other to derive the conclusion. It builds on the erotetic theory of reasoning (ETR) as introduced by Koralus and Mascarenhas (2013).

ETR is an ambitious venture that seeks to redefine and regiment mental model theories by bringing them to the level of formal explicitness that we find in formal semantics. On top of these goals, ETR provides a new insight on what reasoning is about. ETR views reasoning (partly) as a way to answer questions using each premise in a systematic manner to maximize their utility. Premises can serve two purposes in this regard: they can either ask questions or provide elements to answer previously asked questions. Fallacious behavior is going to emerge because of the way premises are used and not because of their content in and of itself. Hence, this account traces the origin of IIFD in reasoning, giving it its name: the reasoning-based account.

The gist of the derivation is the following. Premises containing disjunction-like elements, such as a full-fledged disjunction of course, have the potential to raise alternatives. Such premises ask a question: which of the alternatives is the case?[4]

The first premise P_1 of IIFD $(a \wedge b) \vee c$ is of such a nature: are we a in a $a \wedge b$-situation or c-situation? Each subsequent premise will be used to answer this question. The second premise a overlaps with one the alternatives and not the other. This is taken as a hint that we are in a $a \wedge b$-situation. If this is the case, then the conclusion that b can be made.

This is how people jump to this conclusion, but why? Two reasons. First, reasoning is viewed as a journey towards answering questions. Second, questions here are a set of alternative mental models. One must entertain them during the journey. However, maintaining these alternatives is costly in terms of cognitive resources. A way of resolving this tension is to jump to conclusions even though they might not follow from sound reasoning.

Unlike the original mental models theory, ETR is formulated as a succession of explicit and formal operations. They act upon algebraic formulations of mental models and sentences. The complete system is given in Koralus and Mascarenhas (2013) and Mascarenhas (2014). I will give here a simplified and annotated derivation of illusory

[4]Independent linguistic arguments exist establishing a connection between disjunction and questions. Many natural languages have the same morphemes for the interrogative complementizer and disjunction operator. Malayalam (a Dravidian language spoken in the south of India) is a good example (Jayaseelan, 2004).

(1) John-oo Bill-oo wannu.
John-or Bill-or came
'John or Bill came.'

(2) Mary wannu-oo?
Mary came-or
'Did Mary came?'

(cf. also Japanese 'ka', Korean 'na', several variations of Slavic 'li', Polish 'czy', and so on). In some frameworks (Hamblin, 1958), questions are modeled as sets of propositions, so are disjunctions (Alonso-Ovalle, 2006) and indefinites (Kratzer and Shimoyama, 2002) in many modern approaches to free choice (Aloni, 2007), counterfactuals, exceptional scope-taking. In inquisitive semantics (Mascarenhas, 2009b), disjunctions are at the core of inquisitiveness, they are the building blocks of questions.

inferences from disjunction.

$$\{0\}[\{a \sqcup b, c\}]^{\mathsf{Up}} = \{a \sqcup b, c\}$$
$$[\{a\}]^{\mathsf{Up}} = \{a \sqcup b\}$$
$$[\{b\}]^{\mathsf{MR}} = \{b\}$$

We begin with an empty mental model.
We update with the first premise: the disjunction in it generates two alternatives corresponding to each disjunct.
We update with the second premise: we keep only those alternatives that have something in common with a. This results in the elimination of the second alternative.
Finally, we check if b is an answer.

2.2.2 The interpretation-based account

The interpretation-based account identifies the source of the fallacious behavior in the way premises are interpreted (hence its name). It builds on theories of scalar implicatures as proposed by Sauerland (2004).

On this account, IIFD do not stem from incorrect inference patterns. On the reasoning-based account, the interesting part happens when premises are combined with one another. Their content is used as it is. On the interpretation-based account however, fallacious conclusions follow from perfectly sound and classical reasoning. Nothing interesting takes place at this point. The interesting part happens a step before, when the premises are interpreted. They receive new content from pragmatic enrichment. These theories of pragmatics have proved highly successful and are completely independently motivated. In sum, what looks like a failure of reasoning is instead the result of entirely justifiable interpretive processes.

For the case at hand, scalar implicature is the relevant notion. In a nutshell, the disjunction triggers the strengthening of the first premise P_1 from $(a \wedge b) \vee c$ to $(a \wedge b \wedge \neg c) \vee (c \wedge \neg a \wedge \neg b)$. Taking this modified premise, it is a valid inference to conclude that b is the case from the second premise a.

The philosophy behind scalar implicature is that meaning is derived not only from what is being uttered but also from what could have been uttered but conspicuously wasn't. Thus, when computing the meaning of a sentence, the addressee must take into account both the sounds emitted by the speaker but also a set of alternative sentences that the speaker may have used. In this sense, scalar implicatures are a special kind of inference that uses non-explicit information. Theories differ as to where they attribute the origin of these entailments. Historically, scalar implicatures were first considered as a pragmatic phenomenon derived from Grice's maxims of conversation (Grice, 1975; Horn, 1972). Now there are neo-Gricean approaches (Sauerland, 2004) that incorporate more formalism. Alternative accounts started by Fox (2007) postulate the existence of a syntactic operator whose semantic interpretation reproduces the main result of the pragmatic approach. In what follows I will present derivations in terms of neo-Gricean

approaches but note that syntactic accounts can also derive the relevant strengthened premise. I will hence sometimes use *pragmatic account* as a synonym for interpretation-based account.[5]

An easy gloss of the operations involved in the derivation is accessible using the word *only*. The first premise P_1 is to be understood as only $(a \wedge b) \vee$ only c, which amounts to $(a \wedge b \wedge \neg c) \vee (c \wedge \neg a \wedge \neg b)$. Let's now turn to an extensive description of the strengthening going on. Mascarenhas (2014) provided a thorough discussion of this account. Here I only give the ingredients of the process:

1. find a relevant set Φ of alternatives ϕ to the original sentence. These should be stronger and no more complex than the original sentence.[6] These are the sentences that the speaker could have uttered but chose not to.
2. compute primary implicatures of the set built in 1. They are of the form *it not the case that the speaker believes that* ϕ.
3. compute secondary implicatures, which corresponds to what the speaker believes not to be the case. They are of the form *the speaker believes that not* ϕ, for each ϕ such that the associated secondary implicature combined with the original sentence does not contradict any of the primary implicatures. In other words, ϕ gives rise to a secondary implicature if and only if $\neg\phi$ and P_1 do not entail any member of Φ.
4. conjoin the original sentence with its secondary implicatures to obtain the strengthened meaning.

The crucial alternative to derive the strengthened meaning is $(a \vee b) \wedge c$. It is obtained by a simultaneous substitution of the two logical connectives of the original sentence. Indeed, combining the secondary implicature derived from this alternative with the original sentence yield the correct reinforced premise:

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

which simplifies into

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

[5] I use the word *pragmatic* in a different sense than the one commonly accepted. It usually refers to meaning derivation processes that are not directly encoded in the meaning of words. Instead, these processes are determined by general conversation principles such as Grice's maxims of conversation. Here I use *pragmatic* as a synonym for *interpretation-based* to improve readability. However, I remain agnostic on the account that scalar implicatures should receive (syntactic or properly pragmatic).

[6] Following Katzir (2007), an alternative is said to be no more complex than a sentence if and only if it can be obtained by substituting elements of the sentence or the sub-constituents of the sentence. For instance a is an alternative to $a \vee b$ but $a \vee b \vee c$ is not. The elements substituted must belong to a specified set. For instance *and* and *or* can be substituted but *between* and *or* cannot. Taken together, these two constraints on alternatives generation answer the symmetry problem (Horn, 1972).

2.2.3 Summary

The reasoning-based account and the interpretation-based/pragmatic account both offer a fully explicit and formal derivation of illusory inferences from disjunction as identified by Walsh and Johnson-Laird (2004). The former identifies the source of the fallacy in the way reasoning proceeds upon premises. The latter defends a classical view of reasoning and puts the blame on the way premises are interpreted. Both crucially rely on the presence of a disjunction: for the reasoning-based account, it is the element generating the two mental-models; for the interpretation-based account, it is triggering scalar implicature processing.

These two accounts appear to be in competition, but they needn't be. Without further empirical evidence, four situations are conceivable:

1. both accounts are indeed involved;
2. only the reasoning-based account plays a role;
3. only the interpretation-based account plays a role;
4. none of the accounts I described is correct and we lack the appropriate theoretical tools to comprehend IIFD.

The view defended by the research program launched by Mascarenhas (2014) is that both of them are involved in IIFD. Reasoning and interpretation conspire in driving fallacious conclusions, both making a contribution to a heterogeneous phenomenon.

2.3 50 shades of illusory inferences from "disjunction"

As noted before, on the reasoning-based account, illusory inferences from disjunction depend on the presence of a disjunction-like element that can give rise to alternatives. Disjunctions are not the only construction that possesses this feature. Existential quantifiers and more recently modals (Mascarenhas and Picat, 2019) have also received theoretical accounts that characterize them as generators of alternatives. From this, it follows that we should expect illusory inferences from "disjunction" with these elements. And indeed IIFD are a much broader class than what has been suggested in the previous sections. Here I will review the different versions of IIFD known so far.

The criterion used to guide the discovery of new versions of IIFD ensures that each of them will receive a reasoning-based account. This is because this account is based on alternatives. As long as disjunction-like elements raise alternatives, there will be a reasoning-based account. However, there is no *a priori* guarantee that an interpretation-based account will be available. And indeed there is none for some of them. This allows us to define two classes of IIFD:

- class *A*, which groups IIFD for which there is both a reasoning-based and an interpretation-based account;
- class *B*, which includes IIFD for which there is only a reasoning-based account.

2.3.1 Class *A* illusory inferences

This class contains two versions we have already encountered, characterized by the presence of an explicit disjunction. Problems in this class have an acceptance rate of about 85%.

Propositional case

(9) There is a king and a ten in Kate's hand, or else a queen.
There is a king.
Therefore there is a ten.

It corresponds to the canonical case first discovered by Walsh and Johnson-Laird (2004). Section 2.2 contains a complete description of how fallacious behavior arises in this case.

Universal case

(10) Every boy or every girl came to the party.
John came to the party.
Therefore Bill came to the party.

This is the first example that we encountered back in the introduction. First discovered by Mascarenhas (2014), and studied in detail by Mascarenhas and Koralus (2017) its structure is very much like the propositional case as the disjunction is explicitly realized as such. The similarity gets even stronger if we see a case of ellipsis in the first premise: *Every boy [came to the party] or every girl came to the party*. The difference remaining is the replacement of a conjunction with a universal quantifier. As a matter of fact, universal quantifiers can receive an interpretation in terms of conjunction: they can be seen as a potentially infinitary conjunction over a given domain (here the domain of boys and girls). The first premise has the following logical form:

$$\big(\forall x \in B.P(x)\big) \vee \big(\forall x \in G.P(x)\big) \equiv \\ \big(P(b_1) \wedge P(b_2) \wedge \cdots \wedge P(b_n)\big) \vee \big(P(g_1) \wedge P(g_2) \wedge \cdots \wedge P(g_m)\big)$$

Now that the parallelism is established, deriving the interpretation-based account will be straightforward. The crucial alternative is going to be the one in which conjunction and disjunction are simultaneously replaced.

$$\big(P(b_1) \ \vee \ P(b_2) \ \vee \ \cdots \ \vee \ P(b_n)\big) \ \wedge \ \big(P(g_1) \ \vee \ P(g_2) \ \vee \ \cdots \ \vee \ P(g_m)\big)$$

However, the careful reader will notice that this alternative is not stronger than the literal meaning. The algorithm to compute scalar implicatures proposed by Sauerland (2004)

is then powerless here. Nevertheless, we remark that this alternative is also not weaker. Spector (2007) proposes to include all alternatives that are not weaker than the original sentence. With this revised procedure, the first premise can be strengthened:

$$\Big(\big(P(b_1)\wedge\cdots\wedge P(b_n)\big)\vee\big(P(g_1)\wedge\cdots\wedge P(g_m)\big)\Big)\wedge \\ \neg\Big(\big(P(b_1)\vee\cdots\vee P(b_n)\big)\wedge\big(P(g_1)\vee\cdots\vee P(g_m)\big)\Big)\ ,$$

which simplifies into

$$\big(P(b_1)\wedge\cdots\wedge P(b_n)\wedge\neg P(g_1)\wedge\cdots\wedge\neg P(g_m)\big)\vee \\ \big(P(g_1)\wedge\cdots\wedge P(g_m)\wedge\neg P(b_1)\wedge\cdots\wedge\neg P(b_n)\big)\ ,$$

which we can understand as

$$\big(\forall x\in B.P(x)\wedge\forall x\in G.\neg P(x)\big)\vee\big(\forall x\in G.P(x)\wedge\forall x\in B.\neg P(x)\big)\quad .$$

From this and *John came*, i.e. $P(j)$ with $j\in B$, it follows that Bill came.

An informal gloss of the reasoning-based account is also easy to give: the second premise introduces a boy so that its content overlaps with the first disjunct of the first premise. This constitutes evidence towards this disjunct and drives the fallacious conclusion that every boy, among whom Bill, came to the party.[7]

2.3.2 Class *B* illusory inferences

This class includes IIFD for which there is no pragmatic account. Using the tools of the interpretation-based account yields unwelcome inferences as we will see shortly. These IIFD do not contain an explicit disjunction. They manage to create alternatives through other routes using disjunction-like elements. Problems in this class have an acceptance rate of about 40%. This is below chance but, more importantly, significantly above mistakes on invalid controls. These are mistakes on invalid inferences, which serve as control since we have no theoretical reason to expect them to be compelling fallacies. The performance on invalid controls establishes a baseline for mistakes. Thus, class *B* IIFD cannot be reduced to plain mistakes.[8]

[7] At the time of writing this thesis, the erotetic theory of reasoning is not equipped with adequate formal tools to deal with quantifiers in complete details. Philipp Koralus is currently working on extended ETR to make it able to cope with these cases.

[8] Distinguishing class *B* IIFD from chance is not as straightforward as it may seem. We must resort to high-powered studies involving a large number of participants (Mascarenhas and Koralus, 2015) or resort to more sophisticated experimental designs (Mascarenhas and Picat, 2019).

Existential case

(11) Some pilot writes poems.
John is a pilot.
Therefore John writes poems.

This version was first identified by Mascarenhas and Koralus (2017). The disjunction-like element here is the indefinte that can be analyzed as an existential quantifier. As noted before, a universal quantifier can be seen as a potentially infinitary conjunction. Similarly, an existential quantifier can be seen as a potentially infinitary disjunction (over the set of people).

To see why no pragmatic account is available, let's try to strengthen the first premise. The relevant alternative here is *All pilots write poems*, which we obtain by substituting *all* for *some*. The strengthened meaning is then *Some but not all pilots write poems*. From this and the second premise, it does not follow that John writes poems.[9]

Just as with the universal case, only an informal gloss can be given for the reasoning-based account. The first premise raises the question of which pilot writes poems. The second premise is about a specific pilot. This, in turn, is taken to be a hint towards the fallacious answer that John writes poems.[10]

Modal case

(12) Miranda might be afraid of spiders and play the piano.
Miranda is afraid of spiders.
Therefore Miranda plays the piano.

This version was first identified by Mascarenhas and Picat (2019). Appendix 1 contains the handout of a poster we presented at SALT 29 in May 2019 on this topic.

An account in terms of scalar implicature is not available here. To obtain it, the first premise would have to be strengthened into

$$\Diamond(a \wedge b) \wedge \neg\Diamond(a \wedge \neg b) \Leftrightarrow \Diamond(a \wedge b) \wedge \Box(a \rightarrow b)$$

Besides the fact that this inference is not intuitive. To our knowledge, no account of scalar implicatures derives it.

[9] Except if John is the only pilot in the domain which is not a likely assumption here. Furthermore, notice that the use of *some* is weird when it quantifies over a singleton set: *#some current US president is very tanned*.

[10] A reviewer of the pre-registration suggested that the existential case of illusory inferences from disjunction could be explained solely in terms of pragmatic relevance. This account goes as follows: why would someone say *Some pilot writes poems* and then *John is a pilot* if the two were not relevant i.e. if she did not want her addressee to infer that John writes poems? This story is actually an informal gloss of the derivation proposed by the erotetic theory of reasoning. The question here is not about which theory is best, pragmatic relevance or ETR, but rather which label should we give to the erotetic processes. The debate is mostly a verbal dispute.

On the erotetic theory of reasoning, building up on Ciardelli et al. (2009), we take it that *might(a and b)* raises a single alternative, *(a and b)*. The rest of the derivation follows straightforwardly as a special case of illusory inferences from disjunction.

2.3.3 Summary

The reasoning-based and the interpretation-based/pragmatic accounts have been developed to explain the attractiveness of the propositional case of IIFD. The reasoning-based proved proves its adequacy by successfully predicting the existence of diverse variants of IIFD. The emphasis put on the alternatives in the derivation of the fallacious conclusions predicts that every element that generates alternatives should give rise to IIFD. Some of these new versions lack an interpretation-based account. This feature segregated IIFD into two groups: class *A* (both accounts are available) and class *B* (only the reasoning-based account is available). Empirical evidence comes in support of this distinction. Indeed the acceptance rate of class *A* IIFD is higher than for class *B* IIFD (around 80% vs. around 40%).

2.4 Rationale of the project

For standard accounts of science, the predictive power of the reasoning-based account is a strong argument in its favor. On the other hand, the different acceptance rates of class *A* and class *B* IIFD provide indirect support for the interpretation-based account. This account relies on the idea that the first premise's meaning can be strengthened so that, combined with the second premise, it drives the fallacious conclusion.

(13) P_1: $(a \wedge b) \vee c$
P_1': $(a \wedge b \wedge \neg c) \vee (c \wedge \neg a \wedge \neg b)$

This strengthening involves scalar implicatures. Spector (2007) already suspected that implicatures of this kind were possible. Mascarenhas (2014) built a whole account on the premise that the inference was valid. However, despite the theoretical derivation, to this date, the pragmatic account lacks direct empirical evidence. The work reported here seeks to answer this issue by providing experimental results to support an interpretation-based account of illusory inferences from disjunction.[11] We rely on the following reasoning:

[11] In fact, previous work (Mascarenhas unpublished data) has tried and failed to obtain probative results using this general methodology. A standard technique to block the computation of scalar implicatures is to place the crucial sentence in a downward-entailing environment. In such environments, the direction of logical connections is reversed. Thus, the stronger alternative becomes weaker and scalar implicatures cannot be derived. Classical downward-entailing environments include the antecedent of a conditional and the restrictor of *each*. The syntactic complexity associated with these constructions adds up to the inherent complexity of the first premise of IIFD. They are likely to be at the root of parsing issues for participants. These difficulties could explain the absence of results of previous attempts. For instance, consider *If Jane is brooding and Jeremy is looking at the window, or else Mark is in the garden, then Carry is at the movie theatre.*

1. class *A* and class *B* IIFD differ in terms of whether or not they have an interpretation-based account.

2. there are experimental manipulations that specifically affects pragmatic processes.

3. under such manipulations, differential behavior between the two classes can be attributed to what differs between them, i.e. the interpretation-based account.

4. under such manipulations, differential behaviors between the two classes would constitute empirical evidence in favor of an interpretation-based account of class *A* IIFD.

In other words, interfering with pragmatic processes should affect class *A* IIFD but not class *B*. Indeed, pragmatic processes are thought to be involved only in class *A* IIFD as class *B* IIFD lacks an interpretation-based account. If pragmatic processes are blocked or at least reduced, we expect fewer fallacious conclusions for class *A* IIFD, given that on the two routes toward the mistake has been closed. Class *B* IIFD should not be affected as the closed route was not available to them to begin with.

The goal of the present project is to find reliable and efficient pragmatic manipulations that can be reasonably used with IIFD. The research question underlying this work can be articulated as follow:

What are the effects of pragmatic manipulations on different versions of illusory inferences from disjunction?

In the sections that remain, I will explore two paradigms each implementing two different manipulations. The first one relies on cognitive load. The second one uses priming.

Chapter 3

Reasoning load

De Neys and Schaeken (2007) identified an experimental paradigm that can impair the processing of scalar implicatures. In this section, I will first present this paradigm. Then, I will report our own implementations of it in the study of illusory inferences from disjunction.

3.1 Original study

Using cognitive load, De Neys and Schaeken (2007) successfully decrease the number of scalar implicatures computed by participants. Cognitive load was induced via a dual-task paradigm. Participants alternately faced two tasks so that trials were organized in triplets of the following structure:

- participants were presented a pattern of dots on a 3×3 grid. They were instructed to remember it;
- then they had to assess the truth of a sentence;
- finally, they were asked to reproduce the pattern from before.

In sum, there were two different tasks:

- a memory-load task with patterns to keep in mind while doing
- a truth-value judgment task. Critical sentences were of the form *Some P are Q* where P and Q denoted two groups such that $P \subset Q$, for instance *Some oaks are trees*.

Scalar implicature mechanisms would strengthen these sentences into *Some but not all P are Q* (*Some but not all oaks are trees*). With scalar implicatures, participants should reject the sentence, while they should accept it without. An answer of *false* can be considered to be diagnostic of the computation of a scalar implicature. An answer of *true* signals the absence of such a computation.

When the pattern of dots was complex enough, participants gave significantly more *true* answers than they did when the pattern was simpler (hard condition vs. easy condition). From this result, the authors concluded that cognitive load impairs scalar implicature

processing. Crucially, the responses to control sentences were not affected by the manipulation. This suggests that cognitive load as implemented here affects specifically the computation of scalar implicatures.

The dual-task paradigm seems well suited: it should impair only interpretation-based processes and leave reasoning-based processes untouched. We decided to adapt the paradigm by replacing the truth-value judgment task with an inference-making task. Thus a difference in behavior between class *A* and class *B* IIFD would be attributed to interpretation-based processes, providing an empirical proof of their involvement.

After conducting a series of unsuccessful replications of their results, we revised the methodology of De Neys and Schaeken (2007). In brief, we made the easy condition easier and the hard condition harder. We ran this experiment twice with a different set of inferences for the participants to assess: one in which the focus was on the naturalness of the items and one in which the focus was on their connectedness.

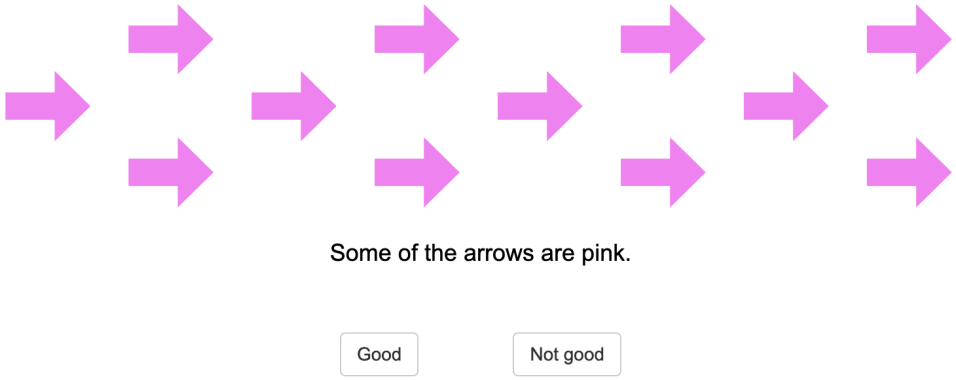


Figure 3.1 – The truth-value judgment task in a successful replication of De Neys and Schaeken (2007)
In our replications, we adapted the truth-value judgment task to make the domain of quantification of *some* explicit. The original experiment had no control on the prior beliefs of participants. The judgments were given out of the blue. This may have contaminated the results. We added a picture about which the sentence was predicating. In this way, we got rid of any concerns regarding prior knowledge of participants.

3.2 A natural version

3.2.1 Methods

Participants

We recruited 123 participants on Amazon Mechanical Turk. We based this decision on two criteria:

- we consulted Paul Marty, researcher at ZAS Berlin, who has ongoing methodological work on the dual-task paradigm as a blocker of scalar implicatures.

- based on pilot data and replications of De Neys and Schaeken (2007), we estimated that this was the figure required to obtain significant results.

Participants had to fulfill the following conditions:

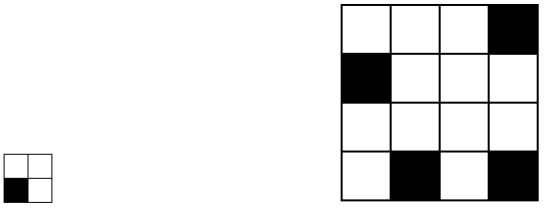
- being located in the US (to maximize the likelihood of being proficient English speakers);
- not being part of the 10% most active workers on the platform (to have normal
- people who would be more likely to run the experiment thoroughly);
- not having taken part in one of our previous experiments (to ensure naiveté vis-à-vis the paradigm).

The mean age was 35.5 (ranging from 18 to 64, $\sigma = 11.5$). 68 participants were female. Participants all received 75 USD cents in compensation for their work.

Procedure and stimuli

We used a dual-task paradigm in a 2×2 within-subjects design. The first task was a memory-load task in which participants had to remember a pattern of black squares on a grid. The second task was an inference-making task in which participants had to decide if a proposed conclusion follows from a set of premises. We manipulated two parameters:

- the difficulty of the memory-load task, by varying the dimensions of the grid and the complexity of the patterns to remember. In the easy condition, a single black square on a 2×2 grid was displayed. In the hard condition, the grid dimensions were then 4×4 and the pattern consisted of either 4 unconnected black squares, either 2 connected and 2 unconnected squares.[1] The patterns were generated on the fly following these rules. See figure 3.2.
- the type of target inferences to assess (class *A* or class *B* IIFD).



(a) Easy condition (b) Hard condition

Figure 3.2 – Examples of grid patterns in the memory-load task

Each trial had the following structure:

- Step 1, the participants were shown a pattern of black squares for 850ms.

[1] Squares are unconnected if they are not in the Moore neighborhood of each other. The Moore neighborhood of a square is the 8 squares directly around it.

- Step 2, they evaluated an inference: they had to decide if a proposed conclusion followed from two premises.
- Step 3, they were invited to reproduce the pattern from step 1 on a blank grid. They were given feedback on their performance on the memory-load task alone.

All trials with a hard pattern (high-load) to remember were presented in a block. Similarly for the trials with an easy pattern (low-load). Following extensive discussion with our Berlin collaborators, we decided to present the hard condition first. In each condition, participants saw 24 inferences. Those items consisted of 16 targets and 8 controls. They were grouped into 2 groups of 24. The group participants saw first was randomly chosen. See table 3.1 for a detailed presentation of the stimuli.

Our prediction was that participants should draw fewer fallacious conclusions in the hard condition compared with the easy condition for class *A* IIFD only. Additionally and in accordance with previous experiments (Mascarenhas and Koralus (2017), unpublished data of Mascarenhas and personal pilot data), we also expected more fallacious conclusions overall for class *A* IIFD than for class *B* IIFD.

Analyses

We excluded the following participants from our analysis:

- those who reported a background consisting in more than one graduate-level course in natural language semantics and pragmatics (to ensure performance would not be contaminated by prior knowledge of the precise goals of the study) (14 participants);
- those who failed to answer correctly to more than 33% of the control inferences (to remove participants who were not focused while doing the task) (31 participants);
- those who reported using notes or diagrams during the task (to ensure that the difficulty of the memory-load task was not artificially altered) (10 participants);
- those who failed to answer more than 50% of the inferences (0 participants).

At the end of the day, we excluded 37 participants.

We analyzed the remaining data (86 participants) using a binomial linear mixed effects model predicting the probability of making a mistake:

- the dependent variable was the answer to target inferences (accepted or not);
- the fixed effects were the condition (hard or easy condition) and the class of IIFD (class *A* or class *B*);
- the by-subject random effects included a random intercept, a random slope for condition, a random slope for class of IIFD and a random slope for the interaction between the two;
- the by-item random effects included a random-intercept and a random slope for condition.

$$\text{ANSW} \sim \text{COND} * \text{CLASS} + (1 + \text{COND} * \text{CLASS} | \text{SUBJECT}) + (1 + \text{COND} | \text{ITEM})$$

Class/Type	Category	Number per condition	Example
A	Propositional	4	It is 55 degrees in Los Angeles and 36 in New York, or else it is 61 in Miami. It is 55 degrees in Los Angeles. *Does it follow that It is 36 degrees in New York?*
	Universal	4	Every linguist or every chemist from Oxford is attending this conference. Alfred, a linguist from Oxford, is attending the conference. *Does it follow that every linguist from Oxford is attending the conference?*
B	Existential	4	Some constitutional lawyer plays tennis. Pat is a consitutional lawyer. *Does it follow that Pat plays tennis?*
	Modal	4	Ray might be a nurse and have a cat. Ray is a nurse. *Does it follow that Ray has a cat?*
Yes-control	*Modus ponens*	2	If yesterday was Wednesday, Carol went to the theater. Yesterday was Wednesday. *Does it follow that Carol went to the theater?*
	Disjunctive syllogism	2	Philippe took a French or a German class. Philippe did not take a French class. *Does it follow that Philippe took a German class?*
No-control	*Modus ponens*	2	If Bruce went to Tokyo, he took a surprising amount of pictures. Bruce stayed home. *Does it follow that Bruce took a surprising amount of pictures?*
	Disjunctive syllogism	2	Clint or Wanda ate the whole cake. Clint loves chamber music. *Does it follow that Kevin ate the whole cake?*

Table 3.1 – Stimuli for the cognitive load experiment. The third column indicates the number of occurences of each category in each condition.

We used Helmert coding and *post hoc* tests to test our predictions with a multivariate t distribution to correct for multiple comparisons. We were interested in the difference between the easy and the hard condition for each class of illusory inferences. To be more precise, we expected:

- a significant difference for class *A*, meaning that fewer illusory inferences were computed in the hard condition, when interpretation-based processes were blocked, compared to the easy condition.
- a non-significant difference for class *B*. Even though an absence of proof is not a proof of absence, those two results taken together would be in line with the conclusion that the manipulation specifically affects class *A* illusory inferences and thus that interpretation-based processes are involved in the generation of class *A* illusory inferences.

The analyses were conducted on RStudio (R Core Team, 2018; RStudio Team, 2016) using the `lme4` (Bates et al., 2015b), `emmeans` (Lenth, 2019) and `afex` (Singmann et al., 2019) libraries.

3.2.2 Results

Memory-load task

The mean percentage of correctly reproduced squares in the hard condition was 71.5% ($\sigma = 15.6$, $\sigma_{\bar{x}} = 1.7$) and 93.8% ($\sigma = 8.2$, $\sigma_{\bar{x}} = 1.0$) in the easy condition. The performance is overall good and confirms the difficulty of the two conditions. This is consistent with previous pilot data, suggesting that participants were appropriately doing the memory-load task.

Inference-making task

The performance on target inferences was similar to previous experiments, with a high acceptance rate for class *A* IIFD, lower for class *B*. Table 3.2 gives the performance on target inferences. The standard deviation is high, pointing to variation between participants, yet the low standard error lends confidence to the estimate of the mean. Furthermore, the model we used takes this variation into account.

Class	Condition	Fallacies in percent	Standard deviation	Standard error
A	Hard	73.8	26.1	2.8
	Easy	79.1	25.3	2.7
B	Hard	34.3	32.5	3.5
	Easy	32.3	32.3	3.5

Table 3.2 – Performance on the target inferences in the natural version

The full model as described above exhibited a singular fit. This means that not all random effects were needed to capture the variance. There is no clear consensus on

the strategy to use to solve singular-fit issues (Bates et al., 2015a; Barr et al., 2013, documentation of the `lmer4` package). We used an algorithm developed by Alexandre Cremers that implements the guidelines proposed by Bates et al. (2015a). This solution simplifies the random effect structure using principal-component analysis and removes correlations between random effects. Table 3.3 reports the statistical details of the analysis.

Class	Condition	Contrast	Estimate	Standard error	df	z-ratio	p-value
A	.	Hard - Easy	0.616	0.178	Inf.	3.468	< 0.01
B	.	Hard - Easy	-0.434	0.203	Inf.	-2.141	0.1036
.	Hard	B - A	2.808	0.409	Inf.	6.858	$<$ 1e$-$4
.	Easy	B - A	3.859	0.499	Inf.	7.731	$<$ 1e$-$4

Table 3.3 – Statistical details of the analysis of the performance on target inferences in the natural version

We detected a significant difference for the acceptance rate of class *A* IIFD between the hard and the easy condition ($z = 3.468$, $p < 0.01$) such that fewer mistakes were made in the hard condition. We did not detect such a difference for class *B* IIFD ($z = -2.141$, $p = 0.1036$). In both condition, class *A* IIFD were significantly more accepted than class *B* IIFD ($z = 6.858$, $p <$ 1e$-$4 for the hard condition and $z = 7.731$, $p <$ e$-$4 for the easy condition). This is congruent with past experiments.

By design, the performance on controls is good, as participants who responded poorly were excluded from the analysis. Table 3.4 gives a summary of the results for controls.

Type	Condition	Fallacies in percent	Standard deviation	Standard error
Yes-control	Hard	93.6	18.5	1.7
	Easy	98.6	17.7	0.6
No-control	Hard	90.8	15.4	2.0
	Easy	89.8	5.9	1.9

Table 3.4 – Performance on the control inferences in the natural version

A *post hoc* model similar to the one described above reveals no effect of the condition on the responses given to yes-controls ($z = -1.726$, $p = 0.245$), or no-controls ($z = 1.929$, $p = 0.164$).[2]

3.2.3 Discussion

We successfully used a dual-task paradigm involving cognitive load to reduce the rate of fallacious conclusions made in response to IIFD. This manipulation only affected IIFD for which a scalar-implicature account, has been proposed. This is the first piece of

[2] ANSW $\sim$ COND $*$ TYPE $+ (1 +$ COND $*$ TYPE$|$SUBJECT$) + (1 +$ COND$|$ITEM$)$ where TYPE can be either yes-control or no-control.

work providing psycholinguistic evidence that supports an interpretation-based account of these fallacies.

Below I respond to a series of objections one might rise to our results.

A possible flooring effect for class B IIFD

A first concern regards the non-significant difference between the easy and the hard condition for class *B* IIFD: a null result being hard to interpret, one could argue that there is actually an effect of our manipulation that we are not able to detect.

This could be due to a flooring effect: the acceptance rate of class *B* IIFD would be too low to be impaired by cognitive load. This is not a crazy assumption as a flooring effect has already been observed with those inferences. ETR predicts an order effect for IIFD: if the second premise is presented first, fewer fallacious conclusions should be drawn (the proof is left to the reader). Experiments easily detect that for class *A* IIFD. However, for class *B* IIFD, a high number of subjects is needed to elicit only a small effect (Mascarenhas and Koralus, 2015). This is attributed to the low acceptance base rate. However, in our case, there is no theoretical argument to suspect a flooring effect for class *B* IIFD. If there were an effect we would expect it to be in the other direction, i.e. increase the rate of fallacies. Let's see why.

ETR is meant to address both the problem of failure and the problem of success of reasoning. The solution to the former has been detailed in section 2.2.1. To account for the later, ETR postulates a costly operation, which is not the default option. This explains several compelling fallacies, among which illusory inferences from disjunction. If cognitive load had any effect on class *B* IIFD, it would be to impair this costly operation, thus increasing the number of fallacious conclusions. This is in line with our results as the non-significant difference between the easy and the hard condition goes in that direction.

If this were the case however our results would be even more interesting. Indeed, there is no reason to suspect that reasoning-based processes would be affected differently in class *A* and *B* IIFD. Thus, if

- we observe a significant difference for class *A* IIFD and
- the manipulation also affects reasoning-based processes, which results in an increase in invalid answers

then this means that the effect of the manipulation on interpretation-based processes overcomes the effect on reasoning-based processes. As both effects go in opposite directions, the contribution of interpretation-based processes would be even larger than what we detected here! In sum, if this remark is valid, if cognitive load also affects reasoning-based processes, then we would have even more confidence in our results.

On significant but small effect sizes

Our results are significant. Yet the effect size is quite small and its standard deviation not negligible. In this sense, our result would be statistically significant but practically

unsignificant.

This is a valid point but it falls beyond the scope of this study. We were interested in the effect of pragmatic manipulations because it was a way for us to test the contribution of interpretation-based processes to illusory inferences from disjunction. What we have shown is that they are indeed involved. In this sense, we reached our objectives by providing the first empirical confirmation of a prediction made by the interpretation-based account.

About connectedness within items

One could object that the items displayed inner connectedness and this alters our results. Let's go back to the logical form of the propositional case of IIFD to see why.

(14) P_1: $(a \wedge b) \vee c$
P_2: a
C: b?

Besides interpretation-based and reasoning-based processes as described above, the conclusion that b is the case can be driven by another third process. It is conceivable that a alone should raise the probability of b, thereby driving the conclusion on its own, irrespective of the IIFD flavor of this reasoning problem. For instance, look at

(15) Some student smokes.
Carl is a student.
Therefore Carl smokes.

The fact that Carl is a student could independently be an argument to conclude that Carl smokes based on some prior knowledge of the habits of students in France. The behavior we observed in our experiment would then be the sum of three distinct processes: interpretation-based and reasoning-based routes, and the connectedness route.

Furthermore, connectedness could be affected by the dual-task paradigm. A story could be that under load, participants are less likely to be able to make connections between sentences a and b. Thus, this route would also be blocked providing another source to the decrease in fallacious conclusions. However, we would expect that to happen in both classes of IIFD. Thus it does not seem a valid objection to our results.

Modulo these few concerns we answered, we successfully provided the first empirical evidence in favor of an interpretation-based account of illusory inferences from disjunction. As a follow-up, we wanted to control for the connectedness issue just discussed. This is the version I report in the next section.

3.3 An unconnected version

3.3.1 Methods

The methods were identical to the previous versions except for two things:

First, we recruited 170 participants. We recruited 50 extra participants compared to the previous version because based on the exclusion criteria presented below, we knew that we would exclude around 30% of our participants. The mean age was 35.1 (ranging from 18 to 64, $\sigma = 11.5$). 93 participants were female. Participants all received 75 USD cents in compensation for their work.

Second, the target inferences were different. They were designed to address the issue of connectedness raised earlier. To select the appropriate materials, we ran a pilot experiment on Amazon Mechanical Turk. We recruited 120 participants and asked them to rate the strength of the connection in a series of sentences. They were presented items of the type *If a then b* and asked to rate them on a scale from 1 (no connection) to 7 (perfect connection). Out of the 59 sentences tested, we selected the best 32 items that had a median of 1, a standard deviation below 1.5 and the lowest mean.

We excluded the following item prior to the analyses.

(16) Ron has done all of his homework or all of his chores.
Ron has solved his math problems.
Therefore Ron has done his English essay.

We find it not to be suited to our purpose and it escaped the proofreading of the stimuli. The problem lies in the indirect connection between the first premise on the one hand and the two following sentences on the other hand. Indeed, *math problems* and an *English essay* do not have to be part of Ron's homework. Thus, this is more a case of indirect illusory inferences such as studies by Sablé-Meyer and Mascarenhas (2019).[3]

Compare with the following item.

(17) Every secretary or every engineer got a raise.
Dolcy, a secretary, got a raise.
Therefore Oliver, a secretary, got a raise.

Here there is no doubt on the link between the first premise and the other sentences.

We excluded the following participants from our analysis:

- those who reported a background consisting in more than one graduate-level course in natural language semantics and pragmatics (to ensure performance would not be contaminated by prior knowledge on the precise goal of the study) (16 participants);

[3]Indirect illusory inferences from disjunction arise when the second premise entails one element from a disjunct of the first premise, i.e. $(a \wedge b) \vee c$, d, b? with $d \rightarrow a$. A complete analysis of them requires a formulation of the erotetic theory of reasoning in terms of confirmation theory.

- those who failed to answer correctly to more than 33% of the control inferences (to remove participants who were not focused while doing the task) (30 participants);
- those who reported using notes or diagrams during the task (to ensure that the difficulty of the memory-load task was not artificially altered) (9 participants);
- those who failed to answer more than 50% of the inferences (0 participants)

At the end of the day, we excluded 39 participants.

The data were analyzed using the same linear mixed effects model as before.

$$\text{ANSW} \sim \text{COND} * \text{CLASS} + (1 + \text{COND} * \text{CLASS} | \text{SUBJECT}) + (1 + \text{COND} | \text{ITEM})$$

3.3.2 Results

Memory-load task

The mean percentage of correctly reproduced squares was 74.9% ($\sigma = 14.0$, $\sigma_{\bar{x}} = 1.2$) in the hard condition and 93.6% ($\sigma = 6.3$, $\sigma_{\bar{x}} = 0.55$) in the easy condition. Again, the performance is overall good and confirms the difficulty of the two conditions.

Inference making task

Table 3.5 gives the performance on target inferences.

Class	Condition	Fallacies in percent	Standard deviation	Standard error
A	Hard	78.7	21.4	1.9
	Easy	82.2	21.9	2.0
B	Hard	38.9	34.3	3.0
	Easy	41.8	35.0	3.1

Table 3.5 – Performance on the target inferences in the unconnected version

The full model presented a singular fit. We applied the same solution as before using Alexandre Cremers's function. Table 3.6 reports statistical details of the analysis.

Class	Condition	Contrast	Estimate	Standard error	df	z-ratio	p-value
A	.	Hard - Easy	-0.006	0.268	Inf.	-0.024	1
B	.	Hard - Easy	0.518	0.223	Inf.	2.325	0.065
.	Hard	B - A	2.997	0.444	Inf.	6.756	$< 1e{-4}$
.	Easy	B - A	3.522	0.631	Inf.	5.582	$< 1e{-4}$

Table 3.6 – Statistical details of the analysis of the performance on target inferences in the unconnected version

We were not able to detect a significant difference between the acceptance rate of the

fallacious conclusions for class *A* IIFD between the hard and the easy condition ($z = 2.325$, $p = 0.065$). As expected, there was no significant difference for class *B* IIFD in between the hard and the easy condition ($z = -0.024$, $p = 1$). In both conditions, class *A* IIFD were significantly more accepted than class *B* IIFD ($z = 6.756$, $p < 1e{-}4$ for the hard condition and $z = 5.582$, $p < 1e{-}4$ for the easy condition). This is congruent with past experiments.

Table 3.7 gives a summary of the performance on controls.

Type	Condition	Fallacies in percent	Standard deviation	Standard error
Yes-control	Hard	95.0	12.9	1.1
	Easy	97.3	10.4	0.9
No-control	Hard	90.3	17.4	1.5
	Easy	93.9	13.2	1.2

Table 3.7 – Performance on the control inferences in the unconnected version

A *post hoc* model similar to the one described above reveals no effect of the condition on the responses given to yes-controls ($z = 1.658$, $p = 0.289$), or no-controls ($z = -1.230$, $p = 0.548$).

3.3.3 Discussion

Our results do not manage to reach the 5% conventional threshold for significance. From this version of the experiment, we cannot conclude that interpretation-based processes are at play in class *A* illusory inferences. This result is in apparent contradiction with the previous section. However, we can put this statement in perspective with three arguments.

An epistemological argument

First, our priors on the role of pragmatic account are high. The baseline differential acceptance rate between class *A* and class *B* IIFD constitutes in and of itself a weak argument (as noted before). The theoretical soundness of the derivation presented before brings further credit to an interpretation-based account of IIFD. Furthermore, this account does not rely on *post hoc* tools designed specifically to handle this case. Quite the contrary, it resorts to scalar implicatures, a concept independently motivated that has proved fruitful to deal with a vast number of phenomena before (free-choice inferences, ignorance inferences, plural, apparent ambiguity of words like quantifiers and disjunction). Thus, even if the likelihood of our data is not in favor of it, this should not be enough to make the cautious reader abandon the idea of an interpretation-based account as a route to illusory inferences from disjunction. Moreover, our priors initially based only on theoretical ground have been substantially increased by the first version of the experiment, which yielded significant results.

Besides, the quality and the necessity of this version of the experiment are subject to

questions.

Less connected but less natural

The quality of our items and their appropriateness for the case at hand are disputable. By putting focus on the unconnectedness in the design of the target inferences, our stimuli are likely to have lost in naturalness. Compare for instance

(18) Tammy works all day long and Dorothy wants a banana, or else Timothy bought plane tickets.
Tammy works all day long.
Therefore Dorothy wants a banana.

and

(19) Sue speaks English and Peter speaks Japanese, or else Jane speaks Spanish.
Sue speaks English.
Therefore Peter speaks Japanese.

Designing items by focusing on unconnectedness led to the selection of unusual sentences. Their combination may result in deviant items, in the sense that they are too artificial. Our participants performed a task that was too outré to expect normal behavior. Thus we have good reason to suspect that our results are not reliable enough to draw firm conclusions and above all to drastically discredit a hypothesis for which we have reasonable priors.

A useless venture

Finally, we can question the importance of making items unconnected. If we consider likely the possibility that fallacious conclusions are driven by connectedness, that does not mean that it is the only force driving the illusory inference. As noted before, the interpretation-based and the reasoning-based accounts are not thought to be exclusive. Each of them makes a contribution to a heterogeneous phenomenon. Connectedness is to be considered as a third potential route towards the error. The key point behind this argument is that this route is *a priori* involved in both classes of IIFD. If we further assume that this route is not affected by the pragmatic manipulation, it follows that connectedness is not something to worry about. This assumption seems reasonable based on *post hoc* analyses.

We ran a model similar to the one above but to analyze the effect of the manipulation in the *modus ponens* yes-controls alone. Indeed they have the same structure as the items we used to test for the connectedness of the propositions making the IIFD. Thus, they may provide some insights about the effect of the manipulation on connectedness processes in acceptance rate.

$$\text{ANSW} \sim \text{COND} + (1 + \text{COND}|\text{SUBJECT}) + (1 + \text{COND}|\text{ITEM})$$

After simplification of the random effect structure that had no effect on the quality of the model ($\chi^2 = 0.546$, df $= 2$, $p = 0.76$), the effect of condition was not significant ($z = 0.671$, $p = 0.502$). Even though an absence of proof is not proof of absence, this gives a weak argument in favor of the following point: the dual-task paradigm does not affect connectedness processes.

Even if it were affected by the manipulation, there is no *a priori* reason to think that it would be affected differently in class *A* and class *B* IIFD. Thus, a difference between those two classes between the two conditions of the memory-load task would still be attributed to interpretation-based processes. Even if connectedness is involved, this will not affect our interpretation of the data.

Thus it follows that constraining the design of our stimuli on connectedness brought no advantage and makes us pay a non-negligible cost in terms of naturalness. This casts serious doubts on the validity of the results of this unconnected version of the experiment. It appears that it was not even necessary in the first place.

A possible way to discard the issue of connectedness once and for all would be to run a modified version of the first experiment. We would use the same material as in the natural version, and prior to the dual-task, we would ask participants to rate the connection between sentences, as we did to select the items to be included in the unconnected version. With these data in hand, we could include the connectedness rating directly into the linear mixed effects model. This would allow to both assess and control the contribution of connectedness to illusory inferences from disjunction.

3.4 Summary

We designed a dual-task paradigm implementing cognitive load specifically targeting interpretation-based processes by reducing the computation of scalar implicatures. We used it to explore the role of these processes in the generation of illusory inferences from disjunction. We observed a significant drop in fallacious conclusions under cognitive load. We take this to be the first empirical evidence in favor of a pragmatic account of IIFD.

Given the goals of this thesis, our mission is accomplished. Still, one may want to strengthen our results by grounding them in more diverse experimental evidence.

As we evoked before, our results may be too tied to the paradigm we use and lack external validity. I gave arguments to support the idea that it is not a fatal objection to our results. Yet, another way to answer it would be to implement cognitive load in different ways. First, we should assess the potential of such paradigms to efficiently and specifically reduce scalar implicatures processing. This would only be minor, and thus easy modifications to the current and De Neys and Schaeken (2007)' experiments: we would need to replace the memory-load task with another task known to induce cognitive load. Among them we may cite:

- n-back tasks (Jaeggi et al., 2010). They also target memory. Participants have to identify if a stimulus is idle to one presented n trials before. These tasks are known to be extremely difficult as n increases. Thus, they may not appropriate.
- Stroop task (MacLeod, 1991). Participants are presented color words written with colored fonts which may or may not match. This defines congruent and incongruent trials respectively. The difficulty is to find the right proportion of congruent and incongruent trials to design an easy and a hard condition different enough. Another shortcoming will be to find a way to measure the performance of the participants to this task. It cannot be grasped with the percentage of correctly reproduced squares only. Both the answer and the reaction time must be considered to judge a participant on a Stroop task.

Another way to give more ground to the interpretation-based account would be to find other ways not revolving around cognitive load to impair pragmatic processes. This is the purpose of the next section.

Chapter 4

Primed reasoning

Bott and Chemla (2016) introduced a priming paradigm in the study of scalar implicatures. Their goal was to explore the mechanisms underlying the pragmatic strengthening of sentences. A seemingly disparate range of phenomena have been analyzed as the product of scalar implicatures. For instance, numerals like *five students* are typically interpreted with an exact meaning, *exactly five students* (Horn, 1972), and *some* is often taken to mean *some but not all* (Horn, 1989). These two phenomena have been argued to be part of a large scalar implicature paradigm. Yet a question remains: is the mechanism of strengthening the same in these phenomena? In other words, if participants are trained to strengthen or not strengthen sentences containing numerals, are they also going to strengthen or not strengthen (respectively) sentences containing *some*?

Bott and Chemla (2016) gave a positive answer to this question. More interestingly to us, yet less surprising, they also show that the priming also works within each member of the paradigm. That is, participants primed to interpret *some* as *some but not all* were more likely to interpret *some* as *some but not all* later in the experiment, but also *five students* as *exactly five students*.

We would like to use a related paradigm to increase or decrease pragmatic processes and observe the effect on the rate of fallacies committed in response to different flavors of illusory inferences from disjunction. However, the interpretation-based account of illusory inferences relies on the enrichment of a sentence containing a disjunction. Strengthening linked to *or* was not investigated by Bott and Chemla (2016). As a first step, we need to ensure the reliability of the priming paradigm regarding disjunction.

I will first present the challenges risen by this first step, how we tried to address them and then discuss some of the shortcomings we encounter along the way.

4.1 Prerequisites

P_1 sentences of the form $(a \wedge b) \vee c$ have three possible readings:

1. $(a \wedge b \wedge \neg c) \vee (c \wedge \neg(a \wedge b))$, a weakly exhaustive reading
2. $(a \wedge b \wedge \neg c) \vee (c \wedge \neg a \wedge \neg b))$, a strongly exhaustive reading

3. $(a \wedge b) \vee c$, an inclusive reading compatible with $a \wedge b \wedge c$

Only the strongly exhaustive reading will give rise to an illusory inference from disjunction on the interpretation-based account. Let's go back to the propositional logical form of IIFD to understand why.

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

The second premise P_2 is compatible with both disjuncts on both the weakly exhaustive and the inclusive reading of P_1. A model in which a, $\neg b$ and c makes the first disjunct true and one in which a, b and $\neg c$ makes the second disjunct true. Taken together, the two premises are compatible with both a model in which b is true and one in which b is false. Thus, we cannot conclude from the premises that b follows. On the strongly exhaustive reading, however, P_2 is compatible only with the first disjunct, thus, b classically follows.

In order to use this paradigm for our purposes, we need to ensure that priming a strongly exhaustive reading of P_1 is possible. The simplest option would be to use the same kind of P_1 sentences as a prime. However, it would more elegant to prime a complex *or* sentence such as P_1 using a simple *or* sentence. They are of the form *a or b* and have two readings:

1. $a \vee b$, an inclusive reading, compatible with $a \wedge b$
2. $(a \wedge \neg b) \vee (b \wedge \neg a)$, an exclusive reading.

The key idea behind this work is that participants primed to adopt an exclusive reading of a simple *or* sentence are more likely to adopt an exhaustive reading of a complex *or* sentences. The question is which one: the weakly exhaustive or the strongly exhaustive one?[1]

4.2 Methods

4.2.1 Participants

We recruited 148 participants on Amazon Mechanical Turk. 96 were women, mean age was 39.7 (ranging from 18 to 64, $\sigma = 12.1$).

Participants had to fulfill the following conditions:

- be located in the US;
- not be part of the 10% most active workers on the platform;
- not taken part in one of our previous experiments.

[1] Conversely, participants primed to adopt an inclusive reading of a simple *or* sentence are more likely to adopt an inclusive reading of a P_1 sentence.

4.2.2 Procedure and stimuli

The experiment consisted of a series of trials in which participants were asked to decide if a sentence was a good match for a picture. They were informed that pictures had been chosen by other people from a limited set and that they had to assess them. There were three distinct types of trials:

- Prime trials. The picture was two Unicode symbols and the accompanying sentence was of the form *There is a* $SYMBOL_1$ *or a* $SYMBOL_2$. Participants possibly received feedback on their decision based on conditions detailed below. This feedback was there to enforce a inclusive or an exclusive reading of simple *or* sentences.
- Probe trials. The picture was three Unicode symbols and the accompanying sentence was of the form *Either there is a* $SYMBOL_1$ *and a* $SYMBOL_2$*, or else there is a* $SYMBOL_3$.
- We had to include biased trials to control for a possible low-level strategy participants might use. Biased trials were similar to prime trials except that the sentence was of the form *There is* $SYMBOL_1$ *and a* $SYMBOL_2$. Because the inclusive reading of simple *or* is compatible with a *and* sentence as noted before, it was possible that participants developed a strategy where they treated *or* as an *and*. Bias trials were here to block this conjunctive strategy. I will discuss this in more details in the discussion.

The different types of possible trials are given in table 4.1.

If participants access an exhaustive reading, only the strongly exhaustive reading of P_1 but not the weakly exhaustive one predicts a different answer to target probe a) and b). More specifically, target probe a) should be rejected only by those who access the strongly exhaustive reading, whereas target probe b) should be rejected irrespective of the reading. Thus:

- asking the question of whether or not we can prime an exhaustive reading of P_1 amounts to asking if participants reject target probe b) more often in the exclusive condition than in the inclusive condition.
- asking the question of whether or not we can prime the strongly exhaustive reading of P_1 amounts to asking if participants reject target probe a) more often in the exclusive condition than in the inclusive condition.

We settled on a 2×2 design. We manipulated:

- the primed reading of simple *or* sentences. This defines the inclusive and the exclusive condition. This factor was between-subjects to maximize the effect of the priming.
- the target probe presented (a) or b)). This factor was within-subjects.

Trials were grouped in triplets consisting of two prime trials and one probe trial. Each of the 6 possible triplets was presented 8 times.[2] Groups of 6 biased trials were presented

[2]Except for target b), which was only presented 4 times and replaced with controls due to a coding mistake that was only detected after the experiment. The statistical analysis we run on the data is robust to

Name	Picture displayed	Sentence displayed
Crucial prime	♦♥	There's a diamond or a heart.
Yes-control prime	♦♣	
No-control prime	♠♣	
Target probe a)	♦♣★	Either there's a diamond and a heart, or else there's a star.
Target probe b)	♦♥★	
Yes-control probe	♦♥♣	
Yes-control probe	♠♣★	
No-control probe	♠♣▲	
No-control probe	♦♣▲	
Yes bias	♦♥	There's a diamond and a heart.
No bias	♦♣	

Table 4.1 – Stimuli for the priming experiment. The exact symbols and colored were genrated on the fly.

at each quarter of the experiment, for a total of 156 trials.

The prime trials included in the triplets were chosen to balance yes and no-answers so that participants could not develop a response strategy:

- in the exclusive condition, each triplet included a crucial prime trial and a yes-control prime trial randomly chosen.
- in the inclusive condition, each triplet included a no-control prime trial on the one hand and a crucial prime trial or a yes-control prime trial (each 50% of the time). This was to further block the conjunctive strategy we mentioned above.

Feedback was given on all crucial prime trials and with a 25% chance on the control prime trials. The order of the prime trials within each triplet was randomized.

No feedback was provided on the probe trials. The controls probe trials were chosen for the same reason, balancing yes and no-answers.

4.2.3 Analyses

We excluded the following participants from our analysis:

- those who reported a background consisting in more than one graduate-level course in natural language semantics and pragmatics (to ensure performance would not be contaminated by prior knowledge on the precise goal of the study) (10 participants);
- those who reported using notes or diagrams during the task (to ensure that participants were not cheating) (29 participants);
- those who failed to answer correctly to more than 33% of the control trial (to remove participants who were not focused while doing the task) (73 participants);

At the end of the day, we excluded 89 participants. This high number is one of our primary concerns regarding this pilot. I will further discuss this point in the discussion.

We analyzed the remaining data (59 participants) using a binomial linear mixed effects model predicting the probability of deciding a good match:

- the dependent variable was the answer to target probe trials (good or bad match between the sentence and the picture);
- the fixed effects were the priming condition (exclusive or inclusive), the target probe trials (target probe a) or b));
- the by-subject random effects included a random intercept, a random- slope for target probe trials;

$$\text{ANSW} \sim \text{COND} * \text{TARGET} + (1 + \text{TARGET}|\text{SUBJECT})$$

We used Helmert coding and *post hoc* tests to test our predictions with a multivariate t distribution to correct for multiple comparisons. The contrasts we were interested in

unbalanced design.

were the difference between the inclusive and the exclusive condition for both target probe. To be more precise, we expected:

- a significant difference for target probe a) between the inclusive and the exclusive condition, meaning that at least the weakly exhaustive reading of P_1 is primed.
- a significant difference for target probe b) between the inclusive and the exclusive condition, meaning that the strongly exhaustive reading of P_1 is primed.

4.3 Results

Performance on the crucial prime trials was very good both in the exclusive condition (94.3% correct answers, $\sigma = 23.2$) and in the inclusive condition (96.0% correct answers, $\sigma = 19.5$). This suggests that the priming was efficient.

Table 4.2 gives the performance on target probe trials.

Target	Condition	No-answers in percent	Standard deviation	Standard error
a)	Inclusive	4.0	19.5	1.1
	Exclusive	18.0	38.5	2.3
b)	Inclusive	3.7	18.8	1.5
	Exclusive	83.1	37.6	3.2

Table 4.2 – Performance on the target probe trials in the priming experiment

The full model as described above converges with no warning. Table 4.3 reports statistical details of the analyses.

Target	Condition	Contrast	Estimate	Standard error	df	z-ratio	p-value
a)	.	Incl. - Excl.	$-$ 2.03	0.739	Inf.	-2.745	< 0.05
b)	.	Incl. - Excl.	-10.14	2.538	Inf.	-3.996	< 0.001
.	Excl.	a) - b)	$-$ 6.55	1.394	Inf.	-4.697	$< 1e{-4}$
.	Incl.	a) - b)	1.57	1.617	Inf.	0.970	0.685

Table 4.3 – Statistical details of the analysis of performance on target probe trials in the priming experiment

We detected a significant difference in the rate of acceptance of target probe a) and b) between the inclusive and the exclusive condition such that both were more rejected in the exclusive condition ($z = -2.745$, $p < 0.05$ and $z = -3.996$, $p < 0.001$).

Performance on all controls trials was at ceiling by design.

4.4 Discussion

We managed to prime exhaustive readings of complex *or* sentences using simple *or* sentences. We primed both the weakly and the strongly exhaustive readings. In this

sense, we extended the results of Bott and Chemla (2016) adding disjunction to the list of primable implicatures.

The results of this pilot are encouraging but the paradigm is not yet ready to be used with illusory inferences from disjunction. Below I review some of the problems we encounter that lead us to adopt the current design and the remaining challenges.

Participants seemed to have adopted low-level strategies in our experiment. Whenever there was a mismatch for one symbol between the sentence and the picture, the rejection rate was high. It is to counter this strategy that the biased trials were added to the experiment. Yet another explanation can be grounded into lower-level strategies.

It is possible that participants paid no attention to the structure of the probe sentences an only focus on the correspondence between symbols displayed and symbols named. This would result in an apparent conjunctive strategy but it also makes another prediction: if it's only about matching symbols and names, we can expect that the more mismatches there are, the more likely participants would be to reject a picture. This was roughly consistent with the mean match rate on probe controls, but no statistical analysis could be conducted to go beyond visual inspection of the data. By following this conjunctive strategy, they could get through the experiment with very little negative feedback. The feedback given on the biased trials should target this relevance strategy. Because participants adopting this strategy would fail on the biased trials, they would be given negative feedback. This, in turn, should make them question their strategy and focus more on the sentences.

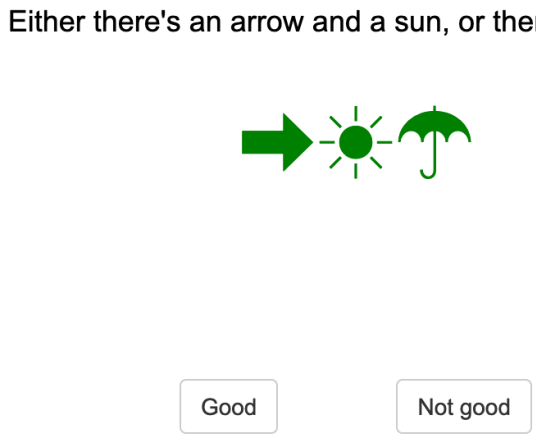


Figure 4.1 – A yes-control probe trial in the priming experiment.
The picture is a good match for the sentence (the first two symbols makes the first disjunct true). Yet, the mismatch between the third symbol displayed and the third symbol named could drive a no-answer if the participant is not paying attention to the structure of the sentence.

These solutions are not efficient enough as we have to exclude almost half of our participants because of mistakes on controls. A possible way to tackle this issue would be to give more feedback to participants. We could give feedback on the control probe trials. This would make participants aware of their mistakes and force them to pay

attention to the structure of the sentences, cutting short any relevance or conjunctive strategies. However, such a modification would make target probe trials stand out of the crowd of trials. This is not something we want, otherwise participants may behave in unpredictable different manners on these trials.

Another solution could be to make the experiment shorter. Amazon Mechanical Turk platform is such that experimenters like us have to specify the estimated length of our experiments. If participants take too long to answer, they cannot get their payment. We set the upper time limit to 25 mn and advertise 20. This estimation was based on our own experience and colleagues'. Participants were way quicker than we were: the median time was 14 mn, the first quartile 11 mn and the third quartile 16 mn. The hurry in which participants seemed to have been in may explain the poor performance on controls.s

A further point to elucidate is to check if priming with numerals or quantifiers is transferable to simple and complex *or* sentences. This is an ancillary goal not necessary to pursue the study of illusory inferences from disjunction with a priming paradigm. Yet, this would give further credit to interpretation-based processes if it was possible to show that blocking pragmatic processes on quantifiers affects the computation of IIFD.

Chapter 5

Conclusion

Psychologists and semanticists have approached reasoning in radically different ways. Semantics, grounded in complex formal frameworks, has focused on sound reasoning. For psychology, on the other hand, reasoning failures are of major interest. Even though they are interested in the same capacity, they worked in quasi-isolation for a long time. Even though they were using linguistic stimuli to conduct their experiments, psychologists rarely sought the expertise of linguists. Semanticists on the other hand mainly ignored reasoning failures and focused on reasoning successes. Both fields ignored important elements that should have constrained the theories they were building to explain their data.

Mascarenhas (2014) proposed to bring together the strengths of the two fields. The goals are two-fold. Adapt pre-existing and develop new formal models grounded in experimental data to explain both failures and successes of reasoning. Use these to better understand how humans reasons and draw the line between interpretive processes and more general-purposed reasoning principles. The present thesis is part of this research program. It sought to provide empirical evidence in favor of a formal account for a diverse reasoning failure: illusory inferences from disjunction.

Two accounts have been proposed for this fallacy. The reasoning-based account identifies the source of the mistakes in the way premises combined. They arise from an urge to answer questions asked by some premises, at the cost of using other premises in non-valid manners. On the other hand, for the interpretation-based account, reasoning operates in perfectly classical ways. Interesting things happen when premises are interpreted: they are strengthened by scalar implicatures. This partitions illusory inferences from disjunction into two classes: class *A* for which both a reasoning-based and an interpretation-based account have been proposed and class *B* for which there is only a reasoning-based account. It is important to note that the two accounts are not competing theories but rather two independent routes to a fallacious conclusion.

We used a dual-task paradigm to specifically impair the processing of scalar implicatures. This resulted in fewer mistakes on class *A* illusory inferences but not on class *B* in which pragmatic processes are not theorized to be involved. We take that to be the first direct empirical argument in favor of an interpretation-based account of illusory inferences from disjunction. The effect size was small but suited to the purposes of the

current study: we were not interested in the potential applications of our results, for instance, to improve reasoning in real life.

We drew the lines of an ambitious priming paradigm for the study of illusory inferences. Its goals are two-fold. First, it will provide a further and independent empirical evidence for the involvement of interpretation-based processes. Hopefully, it will overcome the weakness of the dual-task paradigm in terms of effect size. Thus and second, priming might provide an efficient way to reduce participants' mistakes in a real-life setting.

The methodology we used along this work is also suited to another type of reasoning failures: repugnant validities. These fallacies consist in refusing a yet valid conclusion. Accounts relying on scalar implicatures have been proposed to make sense of them. We could apply the same paradigms to impair pragmatic processes and observe if this results in an increase of logical behavior.

This work provided a further example of the fruitfulness of interdisciplinary approached in cognitive science. Both psychology and semantics can benefit from each other. Building psychologically plausible yet fully explicit models of meaning cannot be achieved without a strong interdisciplinary dialogue. Engaging a discussion between fields that worked apart for a long time is not an easy venture. Here, the spark was a common interest in a phenomenon: illusory inferences from disjunction. They are simple enough to be open to a successful formal analysis, yet they hide a great variety: they can take many different forms and radically different accounts have been proposed for them. In this sense, they are a very adequate tool to understand how reasoning operates and draw the line between interpretive processes and general-purpose reasoning. If, as Ira Noveck said, scalar implicatures are the *drosophilia melanogaster* of semantics, illusory inferences from disjunction might well be the *mus musculus* of reasoning.

Appendix A

Might as a generator of alternatives, the view from reasoning

A.1 Summary

We argue that the epistemic modal *might* is a generator of alternatives in the sense of Hamblin semantics (Kratzer and Shimoyama, 2002) or inquisitive semantics (Ciardelli et al., 2009). Building on methodologies from the psychology of reasoning, we show that *might* patterns with disjunctions and with indefinites in giving rise to a particular kind of illusory inference. The best extant accounts of these illusory inferences crucially involve alternatives, paired with matching strategies (Walsh and Johnson-Laird, 2004) or with question-answer dynamics (Koralus and Mascarenhas, 2013). We present experimental evidence that *might* is a generator of alternatives much like disjunctions and indefinites. We argue that these alternatives have important functions above and beyond those identified in linguistic semantics, as ways of structuring mental representations of information by drawing attention to specific subparts of the representations.

A.2 Background

The semantics of epistemic modals has puzzled linguists and philosophers for decades. Here we address a debate that hasn't been under the spotlight in recent years: the role of *might* as a means of directing hearer attention by generating a single alternative in the sense of Hamblin semantics or inquisitive semantics.

A.2.1 Illusory inferences from disjunction

The erotetic theory of reasoning (ETR) of Koralus and Mascarenhas (2013) holds that reasoning is partly about questions and answers. Some superficially declarative sen-

tences raise issues in the sense of inquisitive semantics (Mascarenhas, 2009a; Groenendijk, 2008), thereby posing questions besides possibly providing some information. Human reasoners entertaining questions look for means of dispelling those questions as swiftly as possible. This desire to reduce the number of alternatives under consideration is responsible for a large class of compelling fallacious inference patterns.

Alternative generators play a central role in this view as they are the elements raising questions. Prototypical examples include disjunction and indefinites. They induce illusory inferences as exemplified below.

(21) John speaks English and Mary speaks French, or else Bill speaks German.
John speaks English.
Therefore Mary speaks French.

(22) Some pilot writes poems.
John is a pilot.
Therefore John writes poems.

Ciardelli et al. (2009) argue that *might* is a generator of alternatives like disjunction or indefinites. Combined with their semantics for *might*, ETR predicts that discourses such as the one below should induce an illusory inference.

(23) Miranda plays the piano.
Miranda might play the piano and be afraid of spiders.
Therefore Miranda is afraid of spiders.

Crucially, ETR does not make the same prediction for the example below, which is plausibly discourse equivalent to the one above but lacks the required question-answer configuration. For notice that, once a has been asserted, it is added to the common ground of the conversation. In a context that guarantees a, $might(a \wedge b)$ and $might(b)$ should have identical effects.

(24) Miranda plays the piano and might be afraid of spiders.
Therefore Miranda is afraid of spiders.

A.3 Study on *might*

A.3.1 Design

We recruited 210 subjects on Amazon Mechanical Turk. 66% of our participants were female. The mean age was 36 (ranging from 18 to 74, $\sigma = 11.4$).

Participants had to solved 18 reasoning problems: 8 targets corresponding to a variation of the illusory inference at stake and 6 controls.

Subjects were randomly assigned one of the four following conditions. Each condition is associated to the question we are interested in and the relevant comparison to answer

it.

Canonical $might(a \wedge b), a \vdash b$

23 Miranda might play the piano and be afraid of spiders.
Miranda plays the piano.
Does it follow that Miranda is afraid of spiders?

- Can *might* trigger illusory inferences from disjunction?
- Comparison with the no-controls.

Reversed $a, might(a \wedge b) \vdash b$

(25) Miranda plays.
Miranda might play the piano and be afraid of spiders.
Does it follow that Miranda is afraid of spiders?

- Is there an order effect?
- Comparison with the Canonical.

P1 $might(a \wedge b) \vdash b$

(26) Miranda might play the piano and be afraid of spiders.
Does it follow that Miranda is afraid of spiders?

- Is this illusory inferences only due to the first premise?
- Comparison with the Canonical and Reversed.

Flat $a \wedge might(b) \vdash b$

iranda plays the piano and might be afraid of spiders.
Does it follow that Miranda is afraid of spiders?

- Is there something erotetic about the fallacy?
- Comparison with the Canonical and Reversed.

Controls were valid and invalid *modus ponens*.

A.3.2 Predictions

We made the following predictions

1. Canonical and reversed (C&R) targets should be accepted significantly more than the baseline for mistakes established by invalid controls.
2. The acceptance of C&R targets should depend on the presence of the second premise, so that premise 1 alone should not explain the fallacy. This means that P1 targets should be lower than C&R targets.
3. Additionally, the plausibly equivalent but "flat" targets should be lower than C&R targets as well.

4. Finally, order effects have been observed with these kinds of illusory inferences (Koralus and Mascarenhas, 2018), readily explained by question-answer dynamics, so we expected canonical targets to be somewhat more attractive than reversed targets.

A.3.3 Results

We analyzed our data using Wilcoxon ranked signed tests. We fixed the threshold for significance at 5%. All of our predictions were borne out except for one. We were not able to exhibit an order effect. Even if C&R are significantly above the baseline for mistakes, their rate of acceptance may not be high enough to detect an order effect. Previous studies has shown that the effect size was small and required bigger samples.

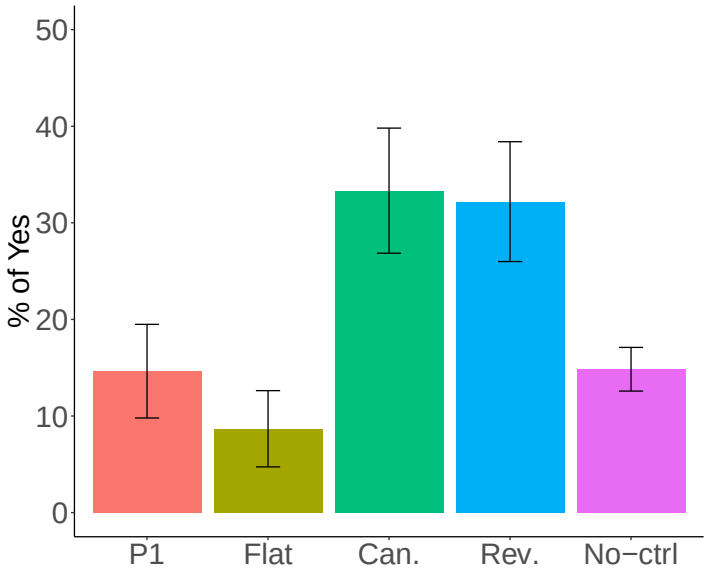


Figure A.1 – Percentage of fallacies committed in response to the different targets

A.4 Different theoretical accounts

A.4.1 Scalar implicatures

An account in terms of scalar implicature is not available here. To obtain it, the first premise would have to be strengthened into

$$\Diamond(a \wedge b) \wedge \neg\Diamond(a \wedge \neg b) \Leftrightarrow \Diamond(a \wedge b) \wedge \Box(a \rightarrow b)$$

Besides the fact that this inference is not intuitive. To our knowledge, no account of scalar implicatures derives it.

A.4.2 Ciardelli and ETR

Ciardelli et al. (2009) argue that the epistemic modal *might* generates alternatives in the relevant sense. In a nutshell, they propose an inquisitive semantics for *might* where $might(\phi)$ is roughly equivalent to $\phi \vee \top$. With their non-classical disjunction, this formula corresponds to an informationally idle but inquisitive meaning that generates two alternatives, one that includes the entire space of possibilities, the other restricted to ϕ.

Feeding this interpretation of *might* into the ETR derives the fallacy. We provide below a simplified version of the derivation.

$$\begin{aligned} \{0\}[\{a \sqcup b, 0\}]^{\mathsf{Up}} &= \{a \sqcup b, 0\} \\ [\{a\}]^{\mathsf{Up}} &= \{a \sqcup b\} \\ [\{b\}]^{\mathsf{MR}} &= \{b\} \end{aligned}$$

Let's gloss what happens at each operation:

1. We start with a blank state. We update with the meaning of $might(a \wedge b)$.
2. Hearing the second premise, we keep only the alternatives that have something in common with a.
3. Finally we check if b is an answer.

A.4.3 Probability account

In a probabilistic framework, asking if a conclusion follows from premises amounts to evaluating the probability of the conclusion given the premises.

In parallel, Lassiter (2016) cites Swanson (2006) who provides a probabilistic semantics for *must*. He extends it to *might*: $might(\phi)$ is to be understood as $P(\phi) > \tau$ where τ is a given threshold.

Putting together this two accounts, it is interesting to compare the analysis one does to understand the canonical case, the reversed case and the flat case.

- **Canonical** $P(b|a \wedge P(a \wedge b) > \tau)$
- **Reversed** $P(b|a \wedge P(a \wedge b) > \tau)$
- **Flat** $P(b|a \wedge P(b) > \theta) > \tau$

Irrespective of the interpretation to give to a probability conditionalized on another probability, one can notice this framework does not predict a difference between the canonical and the reversed case.

A more striking finding is that, as $P(b) \geq P(a \wedge b)$, whenever the condition for the canonical and the reversed case is met, the condition for the flat case is met as well. This means that if one accepts the conclusion for the canonical or the reversed case, one should also accept it for the flat case. Nevertheless, this is not what we observe in our results.

A.4.4 Relational semantics

The facts reported here can be analyzed under a relational semantics for modality , both in its standard guise and with the ordering semantics of Kratzer (1991). The account relies on the following substantive assumptions:

1. when asserting a proposition ϕ, a speaker says, for $w_@$ the actual world, that $w_@ \in \phi$;
2. the world of evaluation is a member of the relevant modal base (reflexivity);
3. there is a crucial existential quantifier over worlds in the lexical entry for *might* that is inquisitive; i.e. it raises a question regarding *which world* we are in;
4. reasoning proceeds as proposed by the erotetic theory of reasoning.

Take the lexical entry for *might* proposed by Kratzer (1991), with the limit assumption for ease of exposition.

$$\textit{might}(\phi) \text{ is true iff there is a } \phi\text{-world among the best ranked worlds}$$

If we interpret this existential quantifier inquisitively, we predict that the following issue would arise "which ϕ world among the best ranked world are we talking about?"

Now take $\phi = a \wedge b$. An assertion of $\textit{might}(a \wedge b)$ will raise the issue "which best-ranked $a \wedge b$-world are we talking about?" Then an assertion of the second premise a, taken to say that $w_@ \in a$, provides the beginning of an answer: the actual world is at least an a-world. The actual world is a member of its own modal base *ex hypothesi*, so one of the possible answers to the issue raised by the first premise is "the actual world is a best-ranked $a \wedge b$-world." By the erotetic mechanisms summarized above, we predict that reasoners should be tempted to conclude that the actual world is indeed the one that answers that question raised by the first premise. When it follows that b is true in the actual world.

A.5 Opening

What about other alternative generators? What about other modals?

A.6 Conclusions

Psychology of reasoning makes extensive use of linguistic stimuli to answer its questions. Semantics can provide invaluable insights to this enterprise. While this is obvious, it is less clear how semantics can benefit from psychology of reasoning. Here we show how the empirical study of reasoning failures is a diagnostic tool semanticists can use to inform their theories. The case at hand is the epistemic modal *might*.

Bibliography

Aloni, Maria (2007). Free choice, modals and imperatives. *Natural Language Semantics*, 15:65–94.

Alonso-Ovalle, Luis (2006). *Disjunction in Alternative Semantics*. Phd diss., UMass Amherst.

Barr, Dale J, Roger Levy, Christoph Scheepers and Harry J Tily (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, 68(3):255–278.

Bates, Douglas, Reinhold Kliegl, Shravan Vasishth and Harald Baayen (2015a). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.

Bates, Douglas, Martin Mächler, Ben Bolker and Steve Walker (2015b). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Bott, Lewis and Emmanuel Chemla (2016). Shared and distinct mechanisms in deriving linguistic enrichment. *Journal of Memory and Language*, 91:117–140.

Cesana-Arlotti, Nicoló, Ana Martín, Ernö Téglás, Liza Vorobyova, Ryszard Cetnarski and Luca L. Bonatti (2018). Precursors of logical reasoning in preverbal human infants. *Science*, 359:1263–1266.

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)*, pages 91–108.

De Neys, Wim and Walter Schaeken (2007). When people are more logical under cognitive load: dual task impact on scalar implicature. *Experimental Psychology*, 54(2):128–133.

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

Grice, Paul (1975). Logic and conversation. In P. Cole and J. Morgan, editors, *Syntax and Semantics: Speech Acts*, volume 3. New York: Academic Press.

Groenendijk, Jeroen (2008). Inquisitive Semantics: Two possibilities for disjunction. ILLC Prepublications PP-2008-26, ILLC, Amsterdam, The Netherlands.

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

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

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

Horn, Laurence (1989). *A Natural History of Negation*. University of Chicago Press.

Jaeggi, Susanne M, Martin Buschkuehl, Walter J Perrig and Beat Meier (2010). The concurrent validity of the n-back task as a working memory measure. *Memory*, 18(4):394–412.

Jayaseelan, K.A. (2004). Comparative morphology of quantifiers. Ms. The English and Foreign Languages University (Hyderabad).

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

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

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

Koralus, Philipp and Salvador Mascarenhas (2018). Illusory inferences in a question-based theory of reasoning. In Ken Turner and Laurence Horn, editors, *Pragmatics, Truth, and Underspecification: Towards an Atlas of Meaning*, volume 34 of *Current Research in the Semantics/Pragmatics Interface*, chapter 10, pages 300–322. Leiden: Brill.

Kratzer, Angelika (1991). Modality. In Arnim von Stechow and Dieter Wunderlich, editors, *Semantics: An International Handbook of Contemporary Research*. Berlin: Walter de Gruyter.

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

Lassiter, Daniel (2016). Must, knowledge, and (in)directness. *Natural Language Semantics*, 24(2):117–163.

Lenth, Russell (2019). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.3.3.

MacLeod, Colin M (1991). Half a century of research on the stroop effect: an integrative review. *Psychological bulletin*, 109(2):163.

Mascarenhas, Salvador (2009a). *Inquisitive Semantics and Logic*. Master's thesis, ILLC.

Mascarenhas, Salvador (2009b). Referential indefinites and choice functions revisited. Unpublished manuscript, NYU.

Mascarenhas, Salvador (2013). An interpretation-based account of illusory inferences from disjunction. Talk given at *Sinn und Bedeutung 18*.

Mascarenhas, Salvador (2014). *Formal Semantics and the Psychology of Reasoning: Building new bridges and investigating interactions*. Ph.D. thesis, New York University.

Mascarenhas, Salvador and Philipp Koralus (2015). Illusory inferences: disjunctions, indefinites, and the erotetic theory of reasoning. In *Proceedings of the 37th Annual Cognitive Science Society Meeting CogSci*.

Mascarenhas, Salvador and Philipp Koralus (2016). Free-form response vs. yes/no-question methodologies in the study of human reasoning. In *38th Annual Cognitive Science Society Meeting CogSci*.

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*.

Mody, Shilpa and Susan Carey (2016). The emergence of reasoning by the disjunctive syllogism in early childhood. *Cognition*, 154:40–48.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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

RStudio Team (2016). *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA.

Sablé-Meyer, Mathias and Salvador Mascarenhas (2019). Assessing the role of matching bias in reasoning with disjunctions. Manuscript under review.

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

Singmann, Henrik, Ben Bolker, Jake Westfall and Frederik Aust (2019). *afex: Analysis of Factorial Experiments*. R package version 0.23-0.

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

Swanson, Eric Peter (2006). *Interactions with context*. Ph.D. thesis, Massachusetts Institute of Technology.

Topál, József, György Gergely, Ádám Miklósi, Ágnes Erdőhegyi and Gergely Csibra (2008). Infants' perseverative search errors are induced by pragmatic misinterpretation. *Science*, 321(5897):1831–1834.

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

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