On the interplay between interpretation and reasoning in compelling fallacies

Léo Picat[1] and Salvador Mascarenhas[2]

[1]Université Paris Cité

UFR de médecine

[2]Institut Jean-Nicod

Département d'Etudes Cognitives

ENS; EHESS; PSL University Paris France; CNRS

Author Note

Declarations of interest: none.

Abstract

We investigate the articulation between domain-general reasoning and interpretive processes in failures of deductive reasoning. We focus on illusory inferences from disjunction-like elements, a broad class of deductive fallacies studied in some detail over the past 15 years. These fallacies have received accounts grounded in reasoning processes, holding that human reasoning diverges from normative standards. A subset of these fallacies however can be analyzed differently: human reasoning is not to blame, instead the premises were interpreted in a non-obvious, yet perfectly predictable and reasonable way. Once we consider these interpretations, the apparent fallacious conclusion is no mistake at all. We give a two-factor account of these fallacies that incorporates both reasoning-based elements and interpretive elements, showing that they are not in real competition. We present novel experimental evidence in favor of our theory. Cognitive load such as induced by a dual-task design is known to hinder the interpretive mechanisms at the core of interpretation-based accounts of the fallacies of interest. In the first experiment of its kind using this paradigm with an inferential task instead of a simpler truth-value-judgment task, we found that the manipulation affected more strongly those illusions where our theory predicts that interpretive processes are at play. We conclude that the best way forward for the field to investigate the elusive line between reasoning and interpretation requires combining theories and methodologies from linguistic semantics and the psychology of reasoning.

Keywords: Deductive reasoning; Pragmatics; Erotetic theory of reasoning; Disjunction; Psycholinguistics

On the interplay between interpretation and reasoning in compelling fallacies

1 Introduction

Failures of human reasoning can in principle stem from two distinct sources. Human reasoning might simply not follow the laws of classical logic. It might for instance resort to non-standard deductive rules (Braine & O'Brien, 1998) or to useful but imperfect heuristics (Tversky & Kahneman, 1974). Alternatively, what look like reasoning failures might instead be the result of perfectly logical reasoning acting upon non-obvious but independently justified interpretations of the premises or the conclusion. In other words, there might be *absolving interpretations* of some apparent reasoning failures, perhaps even many.

We investigate the interplay between reasoning proper and interpretive processes using *illusory inferences from disjunction-like elements*. These are two-premise deductive problems where the first premise sets up a number of competing alternative possibilities.

(1) Mary met every king or every queen of Europe.
Mary met the king of Belgium.
Does it follow that *Mary met the king of Spain*?

The example in (1) is accepted as valid around 75% of the time (Mascarenhas & Koralus, 2017). But it is a fallacy: it is consistent with the premises but not with the conclusion that Mary should have met every queen of Europe, the king of Belgium, and no one else. Extant accounts broadly agree on the process that leads to the mistake (Koralus & Mascarenhas, 2013, 2018; Walsh & Johnson-Laird, 2004). The first premise sets up two possibilities: are we in an every-king situation, or in an every-queen situation? The second premise is directly related to the every-king possibility rather than its alternative, pushing a reasoner to draw the fallacious conclusion.

The crucial driver of fallacious reasoning behind (1) seems to be the disjunction "or" as a generator of alternative possibilities. Accordingly, the literature has established the existence of other fallacies with linguistic elements semantically related to disjunction but superficially very dissimilar, specifically indefinite expressions ("some students") and certain modal expressions ("might") (Bade et al., 2022; Mascarenhas & Koralus, 2017; Mascarenhas & Picat, 2019). These inferences thus form a large and diverse collection of compelling fallacies, while being unified by the presence of a well-studied class of linguistic elements, and being comparatively brief and simple in their syntax.

Crucially, there are alternative accounts of a subset of these illusory inferences that see them as perfectly sound reasoning acting on sophisticated interpretations of the premises. For the case in (1), state-of-the-art theories of pragmatic enrichment predict that the first premise should be interpreted roughly as "Mary met every king *and no one else*, or she met every queen *and no one else*." If this is what the first premise means, then we have an absolving interpretation, for the conclusion *does* follow under this interpretation.

We propose a two-factor theory of these illusory inferences that articulates an account in terms of general-purpose reasoning with one in terms of interpretive processes. In one subclass of these inferences, both processes conspire to create a strong effect, while in the other only one avenue is available, producing an intermediate effect. We provide two sources of evidence for our approach. First, we summarize converging theory and experimental work which is straightforwardly accounted for by a two-factor theory such as ours. Second, in a pre-registered replication of unpublished work (MA thesis by Picat, 2019), we employ a dual-task paradigm from the psycholinguistics literature that is independently known to impair the interpretive processes in question (De Neys & Schaeken, 2007). Confirming our predictions, we found that the rate of fallacious conclusions was significantly more affected in the subclass of illusory inferences where our theory posits two paths to the observed

conclusion. This constitutes the first direct experimental investigation into interpretive processes in these fallacies, and the first successful application of methodologies from psycholinguistics to multiple-premise deductive problems from the reasoning literature.

Zooming out, we argue that only convergent work using state-of-the-art theories and methods from both psychology and linguistics can build the best theories of human reasoning. This is in sharp contrast to the mainstream practice in both fields, which is to merely shift the burden of explanation of mysterious effects to the other field, without properly spelling out *how* precisely the other field can fill the descriptive or explanatory gap, let alone integrating these multiple factors into a complete picture. The theory and experiment we present here, as well as the converging theoretical and experimental results we summarize, instantiate this research program. We show how combining these two fields provides fine-grained insights into a certain class of reasoning failures, illusory inferences from disjunction-like elements.

1.1 Illusory inferences from disjunction

Illusory inferences from disjunction are two-premise deductive problems discovered by Walsh and Johnson-Laird (2004). In (2) we reproduce a representative example of the original version of the fallacy.

(2) 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 *Jane is looking at the TV*?

Concluding that Jane is looking at the TV is fallacious. A situation in which Jane is kneeling by the fire, not looking at the TV, and in which Mark is standing at the window and peering into the garden makes the two premises true but the conclusion false. Yet, this fallacy was accepted 87% of the time (Walsh & Johnson-Laird, 2004).

The particular fallacies discovered by Walsh and Johnson-Laird (2004) can be simplified considerably while keeping their attractiveness and fallaciousness. Consider the inference in (3) below.

(3) Cathy has a dog and Terrence a cat, or else Olivia has a hamster.
Cathy has a dog.
Does it follow that *Terrence has a cat*?

This much simpler case already has all the necessary ingredients from the original example to make it a fallacy. Indeed, in a situation where Cathy has a dog and Olivia a hamster, but Terrence does *not* have a cat, the premises are true but the proposed conclusion false. Accordingly, it is accepted around 77% of the time (this article).

There are two families of accounts of these illusory inferences. Reasoning-based accounts locate the source of the fallacy in the way the premises are combined: humans are not perfect logicians and often combine information in ways that produce fallacious conclusions. On the other hand, interpretation-based accounts assume as a working hypothesis that nothing illogical happens when premises are combined. Rather, humans are reasoning perfectly well but with non-obvious interpretations of the premises, which, once countenanced, show that there was never a mistake to begin with.

In the discussion that follows, it is useful to abstract away from the content of examples as in (3) and to focus on their logical structure. Using italic letters (*a*, *b*, . . .) as symbols for atomic sentences and square brackets ([,]) to indicate scope, we can bring out the structure behind (3) as in (4) below.

(4) [*a* AND *b*] OR *c*
a
Does it follow that *b*?

1.2 Reasoning-based accounts

On a reasoning-based account, the observed inference-making behavior stems from the way the two premises are combined. Premises are interpreted in the straightforward manner, and thus the observed conclusion is fallacious. The only extant reasoning-based accounts of these fallacious inferences come from the mental models paradigm, broadly conceived (Johnson-Laird, 1983). Much if not most contemporary research into deductive reasoning however follows an instantiation of Rational Analysis (Anderson, 1991) in the realm of propositional reasoning sometimes dubbed "The New Paradigm" (Over, 2009), so it is worth explaining in some detail why extant rational-analysis theories and any straightforward extensions of them fail to offer an account of these data.

Sablé-Meyer and Mascarenhas (2021) point out that standard probabilistic accounts predict the observed conclusion *b* for the schema in (4) above to the same extent that they predict an *unobserved* conclusion *c*, under the plausible assumption of independence and flat priors for the atomic propositions *a*, *b*, and *c*. This is because, under the classical-logical interpretations for AND and OR universally presupposed in Bayesian-rationality approaches, the following equalities hold.

$$P([a \text{ AND } b] \text{ OR } c, a) = P([[a \text{ AND } b] \text{ OR } c] \text{ AND } a) = P([a \text{ AND } b] \text{ OR } [a \text{ AND } c])$$

That is, assuming that the three atomic propositions are independent and equiprobable, the conjunction of the two premises puts *b* and *c* on a probabilistic par. It follows that a standard Bayesian-rationality approach cannot distinguish between the attractive conclusion *b* and an altogether unattractive conclusion *c*. This is the case regardless of the particular implementation of such a standard probabilistic account: that is, whether one holds that conclusions in these kinds of problems are accepted if their probability is higher than that of the conjunction of the premises, or instead that conclusions are accepted if their probability conditional on the conjunction of the

premises is higher than a contextually given threshold.

Importantly, none of the foregoing constitutes a general argument against probabilistic approaches, nor does it cast doubt on the insightful idea that decision-making under uncertainty is a core functional aim of the human faculty of reasoning (Oaksford & Chater, 2007). All that this argument shows is that, under plausible assumptions about the relevant prior probabilities, the puzzle exemplified by the schema in (4) is by no means dispelled purely by shifting from classical bivalent logic to the classical probabilistic calculus as the norm. It suggests that there is something either orthogonal to or more fundamental than probabilistic inference at play in this phenomenon. Indeed, Sablé-Meyer and Mascarenhas (2021) argue that in fact a probabilistic mechanism must be part of a complete account of these fallacies, specifically a mechanism assessing *increase in firmness* or *evidential support* (Crupi et al., 2008; Mastropasqua et al., 2010), but that the central representational commitments of some variants of mental model theory regarding disjunction (OR) cannot be done without. In a nutshell, they propose that disjunction offers two possibilities to choose from, and the mechanism whereby reasoners pick an option is that of *(partial) support*: which of the two possibilities raised by the disjunction does the remaining premise provide the most (partial) support for? To simplify discussion and focus on what is specifically revealing about the data at hand in this article, we will refrain from presenting the more complex variants of the account that use probabilistic components, and instead discuss traditional (bivalent) mental-models accounts.

The earliest is given by Walsh and Johnson-Laird (2004), explaining the fallacy as a form of matching bias: reasoners match the second premise *a* to the left disjunct of the first premise *a* AND *b*, leading them to a conclusion of *b*. Koralus and Mascarenhas (2013, 2018) propose a version of mental model theory that incorporates more recent formal and linguistic insights. Their theory brings out most clearly parallelisms between these illusory inferences and other, *prima facie* unrelated ones, so we focus here on their

account.

Koralus and Mascarenhas's Erotetic Theory of Reasoning submits that reasoning consists partly of a question-answer dynamic. Some superficially declarative premises raise questions, which induce pressure on the reasoner to resolve them as soon as possible. Subsequent premises are used in a systematic manner to maximize their utility providing answers to the question under consideration. Fallacious behavior emerges because reasoners will trade off swiftness in dispelling questions under consideration for logical accuracy.

In the case of the illusory inferences of interest, the first premise asks a question by virtue of the disjunction in it. Disjunctions have long been known to share important features with questions. First, many phylogenetically unrelated natural languages show language-internal connections between the morphemes for disjunction and those for interrogation. For example, in Malayalam (Dravidian), the same morpheme "-oo" is used to express an "or" statement in a declarative sentence (5) and to form an interrogative sentence by attaching it to the verb (6) (Jayaseelan, 2001, 2008).

(5) John-oo Bill-oo wannu.
John-**or** Bill-**or** came
"John or Bill came."

(6) Mary wannu-oo.
Mary came-**or**
"Did Mary come?"

This pervasive relationship is not just visible in the morphophonology of natural languages. On entirely independent grounds, natural-language semanticists have been proposing multiple variants of formal approaches to disjunction that coincide with the generally accepted formal approach to interrogatives. In a nutshell, the consensus in semantics and philosophy of language since the work of Hamblin (1958) has been to

analyze questions in terms of answerhood conditions, much like the field analyzes declaratives in terms of truth conditions. The meaning of a declarative is the set of situations in which the declarative is true. Analogously, the meaning of a question is the set of propositions (i.e. declarative meanings) by which the question is answered. Concretely, the meaning of a question as in (6) is the set made up of the two propositions “Mary came” and “Mary did not come.” This approach has spawned an empirically and theoretically rich research field of question semantics and pragmatics.

Interestingly, the idea of question meanings as sets of propositions applies very fruitfully to the formal analysis of disjunction *in declaratives*. In classical approaches to semantics from before the mid 2000’s, a disjunction *a* OR *b* is interpreted in terms of a standard logical connective, either the standard inclusive disjunction (Boolean join) or its exclusive version (Xor). To the best of our knowledge, this classical approach characterizes *all* semantics presupposed by current theories of reasoning with the possible exception of mental-model theories. But in more modern theories of semantics, a disjunction *a* OR *b* gives rise not to a classical-logical meaning, but to a set of *two alternative propositions* (Alonso-Ovalle, 2006; Ciardelli et al., 2019; Groenendijk, 2008; Mascarenhas, 2009). In this view, disjunctions are dramatically different from, say, conjunctions. While *a* AND *b* gives rise to a single proposition, combining the information in *a* with that in *b*, disjunctions have internal structure, and provide later interpretive processes with access to the two individual disjuncts *a* and *b*, rather than only a combination of their content.

Applying these linguistic insights into the meaning of disjunction to our illusory inferences, we expect that the first premise [*a* AND *b*] OR *c* should provide some information, but crucially also *raise a question* with two possible answers, corresponding to each of the disjuncts: are we in an *a*-AND-*b*-situation, or in a *c*-situation? Now the reasoner is entertaining a question, and wishes to dispel it as swiftly as possible by deciding between one of the two possible alternatives. The second premise *a* provides

information that is contained in one of the alternatives (*a* AND *b*) but not the other (*c*). If the reasoner wants to interpret *a* as an answer to the question at hand, they would do well to interpret it as a hint in the *a* AND *b* direction rather than the *c* direction. Accordingly, the reasoner jumps to the conclusion that the question can be answered in the *a* AND *b* direction, whence the observed fallacious conclusion *b* follows. This informal description of the erotetic theory suffices for present purposes, but it is important to note that Koralus and Mascarenhas (2013, 2018) give a complete formal instantiation of these ideas, involving a small number of individually motivated operations on mental representations.

1.3 Interpretation-based accounts

Interpretation-based accounts identify the source of purportedly fallacious behavior in the way premises are interpreted (see for example Hilton, 1995; Stenning & van Lambalgen, 2008). Participants in these experiments are reasoning perfectly well, the experimenter misdiagnosed what happened, for the conclusion *does* follow from the interpretations of the premises that the participants were in fact considering. If there is a failure at all, it was a failure of communication on the experimenter's part, not a failure of the participants' reasoning.

Illusory inferences from disjunction are amenable to interpretation-based accounts of this nature (Mascarenhas, 2013, 2014). The most precise way to articulate such an account is to recruit theories of pragmatic enrichment from the formal pragmatics literature. The version of the theory we outline here is based on the work of Sauerland (2004), Spector (2007), and Chierchia et al. (2012).

These authors give formal implementations and extensions of the standard account of pragmatic enrichment by the maxim of quantity from the work of Grice (1975). In the standard account, listeners assume that speakers are being as informative as required, no more. Chierchia et al. (2012) cash out this idea by proposing that listeners

liberally attach an unpronounced operator with a meaning close to the English word "only" when interpreting utterances from speakers assumed to be cooperative. This approach provides an account of the exclusive interpretations of disjunctions in natural language.

An English sentence like "John or Mary arrived" has as its *literal* meaning an inclusive disjunction, which leaves open the possibility that *both* John and Mary arrived. But in most real-life contexts, an utterance of "John or Mary arrived" will be interpreted *exclusively*: the listener will take it that the speaker believes that John and Mary didn't *both* arrive. We can emulate this stronger meaning of the disjunction by attaching "only" to each disjunct: "Only John or only Mary arrived." The literal interpretation of this last sentence is indeed exclusive: either John but not Mary arrived, or Mary but not John arrived. Modern approaches to pragmatic enrichment postulate hidden operators that perform this duty.

This theory of pragmatic enrichment makes strong predictions for our illusory inferences from disjunction. Recall the first premise of these inferences.

(7) Cathy has a dog and Terrence a cat, or else Olivia has a hamster.

Formal pragmatics predicts that (7) should be interpreted by cooperative listeners in a way analogous to (8).

(8) Only [Cathy has a dog and Terrence has a cat,] or else only [Olivia has a hamster].

In (9) we spell out the truth conditions of this sentence in multiple equivalent ways, in the interest of clarity.

(9) a. It is either only the case that [Cathy has a dog and Terrence has a cat] or it is only the case that [Olivia has a hamster].

b. Either [Cathy has a dog and Terrence has a cat and nothing else that is relevant is the case], or [Olivia has a hamster and nothing else that is relevant is the case].

c. Either [Cathy has a dog and Terrence has a cat **and Olivia doesn't have a hamster**], or [Olivia has a hamster **and Cathy doesn't have a dog and Terrence doesn't have a cat**].

If this is the interpretation of the first premise of illusory inferences from disjunction, then, together with the second premise, the proposed conclusion is no fallacy at all. The second premise adds to the above the proposition that "Cathy has a dog." This contradicts the second disjunct of (9), so the first disjunct must be true, *in classical logic*. From there it follows immediately that "Terrence has a cat," the conclusion that participants in multiple behavioral experiments systematically draw. Crucially, we derived the observed behavior as the product of perfectly sound reasoning processes acting on the interpretations of the premises that state-of-the-art pragmatic theories predict.

It is important to observe that the exact way of cashing out a pragmatic story, Gricean or otherwise, matters greatly. In particular, the classical strategy of interpreting "or" as Xor, the logician's exclusive disjunction, will not suffice. In (10), we give two equivalent formulations of the first premise under an interpretation of "or" as Xor.

(10) a. Either [Cathy has a dog and Terrence has a cat and Olivia doesn't have a hamster], or [Olivia has a hamster and it is not the case that [Cathy has a dog and Terrence has a cat]].

b. Either [Cathy has a dog and Terrence has cat and Olivia doesn't have a hamster], or [Olivia has a hamster and Cathy doesn't have a dog], or [Olivia has a hamster and Terrence doesn't have a cat].

The reasoning from (9) will not work here. There, the first premise had two possibilities,

one of which was directly contradicted by the second premise. But here, as the b. option most clearly shows, there are three logical possibilities. Only one is incompatible with the second premise. Two possibilities are left open, including one in which Terrence doesn't have a cat. This invalidates the conclusion that Terrence has a cat.

To summarize, state-of-the-art theories of pragmatics such as Sauerland (2004), Spector (2007), and Chierchia et al. (2012) provide an account of illusory inferences tracing back the source of the mistake to interpretation and not domain-general reasoning. Such an account is not accessible using classical methods, such as interpreting "or" as Xor.

2 Reasoning *and* interpretation

2.1 The broader paradigm of illusory inferences with disjunction-like elements

Illusory inferences from disjunction stem from the presence of a disjunction. On the reasoning-based account, disjunctions raise alternatives in the form of questions. Pressure to answer such questions produces the observed fallacious conclusions. In linguistic semantics, other elements have been claimed to raise alternatives of a similar kind. This prompts the question whether those elements could drive illusory inferences akin to illusory inferences with disjunction.

Indeed other illusory inferences of this kind exist. First, the structure we presented before can be modulated to create new patterns of illusory inferences from disjunction as exemplified in (11), the universal variant of illusory inferences.

(11) Mary met every king or every queen of Europe.
Mary met the king of Belgium.
Does it follow that *Mary met the king of Spain*?

As in the standard examples, the conclusion in (11) does not follow from the premises. For the following situation makes the two premises true and the conclusion false: Mary

met every queen of Europe, along with the king of Belgium, but no one else. Yet the conclusion is robustly accepted around 75% of the time (Mascarenhas & Koralus, 2017). The parallel with the standard case is salient once one realizes that the universal quantifier ("every") is logically equivalent to an arbitrarily large conjunction of kings, respectively queens. That is, the first premise means roughly "Mary met king *A* and king *B* and . . . ; or else Mary met queen *A* and queen *B* and . . . " This illusory inference is a variant of the standard case rather than a novel pattern in and of itself, since a disjunction is still present, and is presumably the driver of the fallacious conclusion as in the standard case.

The indefinite quantifier "some" constitutes a more radically novel case of illusory inferences. It does not contain an explicit disjunction, making it in form very different from the standard illusory inferences or their universal variant. Yet, indefinite quantifiers like "some" are closely related to disjunction at a mathematical level: they can be seen as arbitrarily large disjunctions. That is, a sentence like "some pilot writes poems" roughly means "pilot A writes poems, or pilot B writes poems, or . . . " These illusory inferences are exemplified in (12).

(12) Some pilot writes poems.
John is a pilot.
Does it follow that *John writes poems*?

The inference in (12) is a fallacy. For it is consistent with the premises for John to be a pilot who does not write poems. Yet, inferences with this structure are accepted around 35% of the time, significantly more than invalid control inferences of a different kind, and enough to be considered an instance of fallacious reasoning that requires an explanation (Mascarenhas & Koralus, 2017). Despite the lower attractiveness of these fallacies, compared to the disjunctive examples discussed above (35% vs. 75%), we will apply the label "illusory inferences" to these cases as well, so as to highlight their structural

proximity with disjunctive cases. We will address the issue of the two classes of fallacies' sharply different acceptance rates head-on shortly.

Mascarenhas and Picat (2019) recently discovered yet another new case of illusory inferences, replicated and studied in greater depth by Bade et al. (2022). Ciardelli et al. (2009) argued that the epistemic modal "might" also has a disjunction-like semantics. While disjunction itself and indefinite quantifiers introduce two or more alternatives, "might" would put forth a single alternative for consideration. On the erotetic theory, this would raise the question whether this alternative is the case. Accordingly, the erotetic theory expects "might" to drive fallacious conclusions in the right configurations. Building on this theoretical work, Mascarenhas and Picat (2019) proved that these modal constructions bring about illusory inferences, as exemplified in (13).

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

Similarly to the indefinite case, these inferences are accepted around 35% of the time, significantly more than mistakes are made on invalid controls (Bade et al., 2022; Mascarenhas & Picat, 2019). Yet, again, the conclusion does not follow. All we have to assume is that it is also possible that Miranda plays the piano and isn't afraid of spiders, and we can model the premises while falsifying the conclusion.

To summarize, the standard case of illusory inferences is only the tip of a larger iceberg that later studies have been investigating. This rich paradigm of superficially very dissimilar inference patterns is unified by the presence of linguistic elements in the first premise that have independently been argued to generate alternatives and trigger question-like interpretations.

2.2 Two classes of illusions

In view of this unifying property, there is a puzzling difference in acceptance rates between the standard case and its universal variant on the one hand, and the indefinite and the modal case on the other (75% vs. 35%).

The standard illusory inferences from disjunction are explained by both the reasoning-based and interpretation-based accounts, as reviewed earlier. The same goes for the universal variant, since it contains a disjunction and is therefore structurally identical to the standard case for all relevant purposes. What of the new more exotic cases with indefinites and modals?

First, the reasoning-based route. We saw that, from a logical perspective, indefinites are like an arbitrarily large disjunction. We also saw that disjunctions are like questions. Accordingly, theories from natural language semantics hold that an indefinite construction like “some pilot writes poems” raises the same alternatives as the question “which pilot writes poems?” (Ciardelli et al., 2019; Kratzer & Shimoyama, 2002/2017). On the erotetic theory of reasoning, and just like the standard case, this question induces pressure in the reasoner’s mind to resolve it as swiftly as possible. The second premise provides a hint of an answer by asserting that John is a pilot. The theory now predicts that the reasoner will jump to the fallacious conclusion that John is the pilot who writes poems in question. Mascarenhas and Picat (2019) give an explanation of the modal case of illusory inferences along the same lines. This constitutes a reasoning-based account for both the indefinite and modal cases of illusory inferences.

The interpretation-based account however is not available for the indefinite and modal cases. Independently motivated pragmatic theories act to strengthen the first premise of the standard case so that once combined with the second premise the conclusion follows logically. But in the indefinite case, the same theories predict a strengthening for the first premise that does *not* guarantee the validity of the conclusion. Specifically, these theories predict that the first premise of the indefinite inference should

be interpreted as “some but not all pilots write poems.” From there, learning from the second premise that “John is a pilot” is nowhere near enough to logically conclude that “John writes poems.” Similarly for the modal case (Mascarenhas & Picat, 2019). Consequently, the pragmatic theories that provided an absolving interpretation for the standard illusory inference are of no help with indefinite and modal illusions.

To sum up, there are two kinds of illusory inferences. On the one hand, we have the standard case and its universal variant, with high acceptance rates (75%), and two ostensibly competing accounts, one based on reasoning and the other on interpretation. On the other hand, the indefinite and modal illusions, with intermediate acceptance rates (35%), and only one possible account, a reasoning-based account in terms of the erotetic theory or mental model theory.

We propose that, in the standard case and universal variant, the two possible accounts are not in competition, but instead constitute two paths leading to the same inference-making behavior. In these *dual-source* illusory inferences, the reasoning-based and interpretation-based accounts conspire to create a very strong effect. By contrast, in *single-source* illusory inferences as produced by indefinites and modals, only one path leads to a fallacious conclusion, the reasoning-based path, predicting the observed intermediate acceptance rates.

The diverging acceptance rates constitute only indirect evidence in favor of this two-paths theory of illusory inferences. A more direct empirical strategy would be to interfere with one or the other source of fallacious behavior.

Sablé-Meyer and Mascarenhas (2019) introduce a means of blocking the pragmatic interpretive processes at play in the standard disjunctive case.

(14) Everyone at this party is either French or a linguist.
That guy is European.
Does it follow that *that guy is French*?

Participants accepted this inference around 40% of the time, a rate comparable to that of single-source illusory inferences. Since this example contains a disjunction and not an indefinite or a weak modal, *why* do we see the lower acceptance rates characteristic of indefinites and weak modals, and not the ceiling acceptance studies have repeatedly found with disjunction?

In (14), the expected pragmatic strengthening of the first premise is “Everyone at this party is either French and not a linguist, or a linguist and not French.” Together with the second premise “That guy is European,” it does *not* follow that “That guy is French,” since it is compatible with the premises but not the conclusion that that guy be say a Spanish linguist. That is, unlike the standard disjunction case, where the pragmatic strengthening of the first premise together with the second premise validated the proposed conclusion, in (14) there is no such absolving interpretation for the conclusion “That guy is French.” This is a fallacious conclusion on any of the normative standards we’ve considered here (deductively as just demonstrated, probabilistically under the plausible assumption that there aren’t necessarily more French individuals than linguists at the party). By contrast, the reasoning-based account of this inference is just as operative in (14) as in the standard case discussed at length above. Specifically, the first premise of (14) induces the erotetic state “Are we in an everybody-is-French situation or an everybody-is-a-linguist situation?” whereupon the second premise, overlapping in content with the first option “French” (for it is entailed by it) but not the second option “linguist,” prompts reasoners to answer the question in the “everybody-is-French” direction, predicting the fallacious conclusion that “That guy is French.”

In our terms in this article, this variant of the standard disjunctive case altogether lacks an interpretation-based path to the conclusion while maintaining a reasoning-based path, and accordingly it has an intermediate acceptance rate. This provides tantalizing evidence in favor of our two-factor theory, in that we can incorporate this data point and unify it with the standard disjunctive case within one and the same two-factor analysis.

But this comes at a cost: the transformation from the standard case to something of the shape of (14) is much too heavy handed. Not only is a universal quantifier introduced in subject position, the conjunction in the first disjunct is replaced here by a non-composite predicate “French,” and therefore the relation between the first disjunct and the second premise is at the very least syntactically very different in the two cases. In the standard disjunctive case, we have a conjunction (*a* AND *b*) which entails one of its conjuncts (*a*). Here, we have a subset (French) which entails its superset (European). Ideally then, we would employ an experimental manipulation that leaves the original standard disjunctive case *completely unchanged*, to show in the most direct fashion possible that the high acceptance rates of those cases are as we propose the convergence of two very different mental processes leading to the same inference-making behavior.

Chung et al. (2022) make important progress in a related direction, introducing a novel paradigm to study these kinds of inferences by factoring out the role of language as much as possible. They report a version of the standard case of illusory inferences where the crucial two premises were presented without language, purely on the basis of visual animations. In the context of a water-based mechanism involving a system of pipes and gates that block or allow the flow of water, they constructed a visual version of standard disjunctive illusory inferences, instantiating the familiar schema in (4), repeated below.

(15) [*a* AND *b*] OR *c*

a

Does it follow that *b*?

Participants saw a water circuit with one entrance at the top which splits into two branches that later merge into one exit. One of the branches has two gates in succession (call them *a* and *b*), while the other only has one (call it *c*). Initially these gates are all closed, blocking the flow of water. The part of the mechanism involving the gates is then

concealed, and participants observe that water starts flowing at the bottom. At this point, if they are following the plot, they are likely to represent the proposition that "Either gates *a* and *b* have opened, or gate *c* has." Then the occluder gets partially removed, to reveal that gate *a* is now open, while still concealing the states of gates *b* and *c*.

About 30% of the time, participants concluded that gate *b* must have been open too, significantly above a baseline measure of bias in favor of *b*, and significantly above the acceptance rate of invalid control problems of different structure. A conclusion of *b* is precisely the fallacy of interest here, and since the premises in this study aren't conveyed linguistically, interpretation-based processes cannot be invoked to absolve the fallacious conclusion. The only path to the fallacious conclusion relies on reasoning-based processes. Accordingly, the acceptance rate is much closer to that of single-source illusory inferences than dual-source ones.

These two very different studies constitute converging evidence that indeed (A) some of these illusory inferences can be produced by at least two mechanisms, and (B) the reasoning-based theory we discussed above in terms of question-like mental representations and question-answer dynamics does not depend on *language* proper, but instead says something general about the human representational arsenal. But these studies dramatically change the mode of presentation of the original language-based schema, and are therefore only indirect evidence for our approach. If all of the foregoing is on the right track, then it should also be possible to devise a manipulation that can interfere with the linguistic version of dual-source inferences, crucially without changing the way information is conveyed, hindering the interpretive processes that constitute one of the avenues for the observed inference-making behavior.

Every theory of pragmatics already predicts that certain linguistic contexts decrease the frequency of the kind of strengthening processes crucial to the interpretation-based account. For example, from the sentence "John or Mary arrived," the listener will take it that only one arrived, not both. But in the sentence "if John or

Mary arrived, Bill will be surprised," this exclusivity inference disappears: in a situation where John *and* Mary arrived, we expect Bill to be surprised on the basis of the sentence at hand. Thus, an if-clause is a linguistic element well-known to reduce pragmatic processes (Gazdar, 1979; Horn, 1972). Embedding illusory inferences in if-clauses should impair the interpretation-based path to dual-source illusory inferences, reducing fallacious conclusions. The problem with such a strategy is the extraordinary complexity of the stimuli it would involve. The illusions of interest are already made up of multiple sentences, to embed them in even more complex constructions would come at too hefty a price in parsing.

The psycholinguistics literature provides subtler ploys to interfere with interpretive processes. De Neys and Schaeken (2007) successfully used a dual-task paradigm to hinder the very kinds of pragmatic processes at stake in our interpretation-based account of illusory inferences. They combined a memory-load task aimed at saturating the working memory of participants, and a truth-value judgment task. Participants had to evaluate the truth of sentences displaying the same pragmatic enrichment at play in dual-source illusory inferences. Under cognitive load, participants accessed fewer strengthened interpretations of these sentences. Marty and Chemla (2013) and Marty et al. (2013) used a similar paradigm to further investigate and characterize the mechanisms at play in pragmatic strengthening.

In view of the multiple successful implementations of the dual-task methodology, we chose to use this paradigm. We implemented cognitive load to interfere with the pragmatic processes presumably at play in dual-source illusory inferences.

3 Experiment

We preregistered the methods and full analysis script of our experiment (Picat & Mascarenhas, 2020a). All of our materials and data are available online (Picat & Mascarenhas, 2020b). To ensure maximum reproducibility, we also published the code

used to run the experiment (Picat & Mascarenhas, 2020b). The present experiment can itself be considered as a replication of work conducted by Picat (2019) in his master's thesis. We build on the same experimental paradigm and improve it in a few ways.

3.1 Methods

3.1.1 Participants

We set out to recruit 150 participants on Prolific, a crowd-sourcing platform based in the United Kingdom. The studies we have cited using a dual-task paradigm to affect pragmatic processes used samples between 16 and 60 participants. But the literature on illusory inferences used samples ranging from 120 to 210 to diagnose the more subtle single-source illusory inferences. So we opted for 150 participants, which had also been used with promising results in the earlier incarnation of this study (Picat, 2019). Three participants did not submit their responses to our experiment, so we effectively recruited 147 participants. Informed consent was obtained for each of the participants.

Participants had to be native speakers of English and not have taken part in one of our previous experiments on this platform. The mean age was 33.74 (ranging from 20 to 70, SD = 12.18). Seventy percent of participants were female. All participants received GBP £2.50 (£7.50 per hour) in compensation for their work.

3.1.2 Procedure and materials

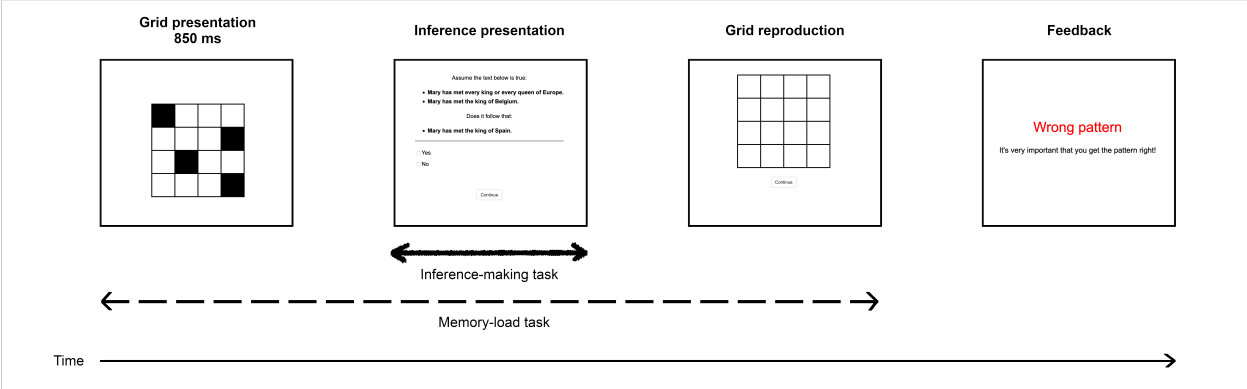
The experiment presented itself as a web page written using the `JsPsych` library (De Leeuw, 2015) and custom plugins developed in our lab.

We used a dual-task paradigm in a 2×2 within-subjects design. Participants had to perform two tasks, one embedded in the other. The outside task was a memory-load task, in which participants had to remember and reproduce a pattern of black squares on a grid. The embedded task was to assess if a proposed conclusion followed from a given set of premises (inference-making task).

Each trial in our experiment had the same structure, as follows. For 850ms, participants saw a randomly generated fresh pattern of black squares on a grid. Then, participants had to evaluate an inference. They could answer “yes” (the conclusion follows), “no,” or choose not to answer. Then, participants had to reproduce the pattern from before on a blank grid, with no time limit. They received immediate feedback on their performance in the memory-load task alone (“correct” if they reproduced 100% of the squares correctly, “almost correct” if 50% or more, “wrong” otherwise). The structure of a trial is illustrated in Figure 1.

Figure 1

Structure of a trial

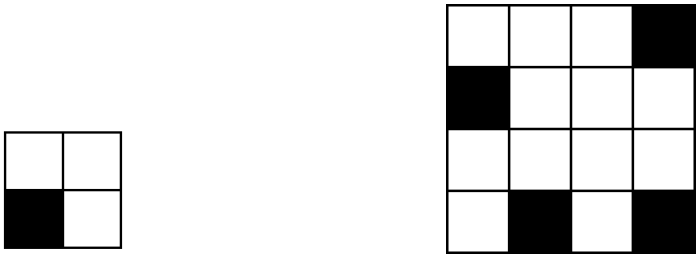


We manipulated two parameters within subjects: the difficulty of the memory-load task and the class of inference to assess. The difficulty of the memory-load task had two levels. In the easy condition, there was a single black square on a 2×2 grid. In the hard condition, there were four black squares on a 4×4 grid. The patterns were such that no more than two squares were ever adjacent. Figure 2 shows a sample pattern for each condition.

The two classes of target inferences participants assessed were dual-source and single-source illusory inferences, as summarized in Table 1. An anonymous reviewer raised an interesting concern about the dual-source targets we label “universal,” which deserves discussion. The idea is that the first premise, with two universal quantifiers

Figure 2

Examples of grid patterns in the memory-load task



(a) *Easy condition* (b) *Hard condition*

("every"), might inadvertently be parsed by our participants by leaving out the second universal. Consider for concreteness the instantiation of the first premise in (16a), with the putative alternative parsing of concern in (16b). We highlight the quantifiers of interest.

(16) a. Mary has met **every** king or **every** queen of Europe.

b. Mary has met **every** king or queen of Europe.

The first thing to observe is that, if participants are parsing (16a) as (16b), then the conclusion that, for any king of Europe, Mary has met that king, is in fact *valid*. This is because "every [*A* or *B*]" is equivalent to "[every *A*] and [every *B*]," in the normative standards we consider in this article (chiefly deductive but also probabilistic validity). We think it is plausible that some participants might parse the sentence in this way, but we submit that even in that case this class of stimuli at worst introduce a *conservative* confound into our experiment. We are aware of no grammatical mechanism of English whereby (16a) *canonically* has the interpretation in (16b). This means that (16b) must be a misparse of (16a), the interpreter simply skipped or otherwise ignored the second quantifier, they must have failed to attend to it.[1] But then the probability of misparsing

[1] At the request of an anonymous reviewer we conducted an ancillary experiment to compare the truth

Table 1

Types and logical schemata of target inferences

Single source		**Dual source**	
Type	Schema	Type	Schema
modal	might [*A* and *B*]	disjunctive	[*A* and *B*], or *C*
	A		*A*
	B?		*B*?
indefinite	some *A B*	universal	every *A* or every *B*
	x A		*x A*
	x B?		*y A*?

(16a) in this way must *go up* as a function of cognitive load, not down. Consequently, if some participants are treating sentences like (16a) as though they were of the shape of (16b), we would expect them to double down on this unusual parse in the heavy-load condition, not correct it. Since the hypothesis of interest in this study is that the proposed conclusion in these inferences should be accepted *less often* under heavy cognitive load than under light or no cognitive load, this potential confound has no real risk of introducing a false positive, only at worst a false negative. We return to this question in

---

conditions of our universal variant of the illusory inferences with disjunction and a related version substituting “either-or” for plain “or” and “all” for “every” (“Mary has met either all kings or all queens of Europe”). The two kinds of sentences are expected to have same truth conditions, and accordingly we found that they are interpreted in very much the same way. The only difference we found was clearly related to the misparse under discussion: in a scenario where the correctly parsed sentences were true but the misparse in (16b) was false (Mary met all kings but no queens), participants were at ceiling in their judgment that the “either-or all” variant was true, but were slightly less likely to judge that (16a) was true. This is exactly as expected, since the version with “either-or all” does not have this plausible misparse. Importantly, this supporting study allows us to conclude with greater confidence that the misparse in (16b) is quite rare.

the discussion of our results.

Participants also assessed a fixed collection of control inferences, of a different nature, not predicted by any extant theory to produce fallacies. Valid controls were modus ponens inferences and disjunctive syllogisms. Invalid controls were pseudo-modus ponens inferences of the form IF *a* THEN *b*; *c*; *b*? and pseudo-disjunctive syllogisms of the form *a* OR *b*; NOT *a*; *c*? or *a* OR *b*; *c*; *b*?. We give the complete list of our stimuli in the supplementary materials.

We presented all trials with a hard pattern to remember in a single block (high memory load). Similarly for the trials with an easy pattern (low memory load). The hard block was presented first. We made this choice based on pilot studies trying to replicate De Neys and Schaeken (2007). The replication was successful only if the hard block came first. This choice also ensured the difficulty of the memory-load task. The opposite order could have made participants too familiar and at ease with the task. This would have diminished the difficulty and thus the difference between the two conditions. In each condition, participants saw 24 inferences: a balanced set of 8 controls and 16 targets.

The instructions included examples of the two tasks to introduce the participants to the technical aspects of the experiment. We insisted that they should not use notes or diagrams. We explained and exemplified the concepts of valid and invalid conclusion. In (17) we reproduce an example of an invalid conclusion and the explanation provided to participants as to why it is invalid.

(17) **Mary is at home**.
If Mary is at home and Ann is playing tennis, then Bill is at work.
Does it follow that *Bill is at work*?
No. (Because the first two sentences could be true and the conclusion false. If Mary is at home but Ann is not playing tennis, the boldface sentences are true, but the proposed conclusion is false.)

Before the core of the experiment started, participants answered 6 training trials.

3.1.3 Predictions

Our central prediction is that participants should fall for fewer dual-source illusory inferences in the hard condition compared to the easy condition. This is because by hypothesis interpretation-based processes are involved only in dual-source illusory inferences, and the hard condition interferes with interpretation-based processes. Crucially, the difference in acceptance rate between the hard and the easy condition should be higher for dual-source illusory inferences than for single-source illusory inferences. This is because by hypothesis interpretation-based processes are not involved in single-source illusory inferences.

Participants should draw more fallacious conclusions from illusory inferences than they make mistakes on invalid controls, no matter the difficulty of the memory-load task. Invalid controls serve as a baseline for uninformative reasoning mistakes. If participants answer “yes, it follows” more often to illusory inferences than to invalid controls, then one cannot attribute these extra errors to noise alone. As is standard practice in the reasoning literature, we take this to be a diagnostic tool to detect fallacies.

We are agnostic about the possible effects of the memory-load task on the reasoning processes in question, from erotetic theory or mental model theory. These theories derive fallacious behavior in these fallacies from fast and automatic mental processes that should have very low cost. This suggests that cognitive load might have only a small to inappreciable effect on such mechanisms. Additionally, on the erotetic theory, the process that uses incoming new information to select one of the possible answers to the question under consideration should in fact free up working memory whenever it is successful, by reducing the number of alternative possibilities under attention.

Despite these considerations, we refrain from assuming that memory load has no

effect on the relevant reasoning processes. We do however take it that, whatever the effect on reasoning processes, it should be the same for dual-source and for single-source illusory inferences. This is because the same reasoning processes are involved in the two classes of inferences, both on the erotetic theory and on mental model theory.

3.1.4 Statistical analyses

All analyses were conducted in `R` (R Core Team, 2018) with RStudio (RStudio Team, 2016) using the `tidyverse` (Wickham et al., 2019), `lme4` (Bates et al., 2015), `rstan` (Stan Development Team, 2023), `rstanarm` (Goodrich et al., 2022), `rstantools` (Gabry et al., 2020) and `loo` (Vehtari et al., 2022) libraries.

3.2 Results and discussion

Following the criteria set out in the pre-registration, we excluded three participants who reported a background consisting in more than one graduate-level course in natural language semantics and pragmatics; six participants who reported using notes or diagrams during the tasks despite our instructions; five participants who failed to answer correctly to more than one third of the valid controls. In total, we excluded 14 participants and analyzed the responses of 133 participants.

We analyzed our data through model comparison of binomial linear mixed-effects models (Bates et al., 2015; RStudio Team, 2016) predicting mistakes (yes responses) as a function of class of inference and difficulty of the memory-load task. We included all relevant random effects, as long as the models converged, as recommended by Barr et al. (2013). We used nested linear model comparisons on relevant subsets of our data to test our predictions. We used Holm correction for multiple comparisons. Our analysis was pre-registered (Picat & Mascarenhas, 2020a).

3.2.1 Memory-load task

In the hard condition, participants reproduced 71.7% of the squares correctly (SD = 14.6, SE = 1.3) against 94.1% in the easy condition (SD = 7.2, SE = 0.6). This attests to the difficulty gap between the two conditions. Yet, performance is overall good and indicates participants were appropriately performing the memory-load task.

3.2.2 Inference-making task

Table 2

Performance on target inferences

Class	Difficulty	Fallacies in percent	Standard deviation	Standard error
Dual-source	Hard	75.6	25.4	2.2
	Easy	78.7	24.5	2.1
Single-source	Hard	34.1	30.6	2.7
	Easy	32.9	32.2	2.8

Table 2 gives a summary of the results on illusory inferences. All of our predictions were borne out. First, we looked at our central prediction that cognitive load should affect dual-source illusory inferences to a greater extent than single-source illusory inferences. We fitted participants' responses to illusory inferences into a binomial linear mixed-effects model, predicting the answer to illusory inferences from the difficulty of the memory-load task, the class of illusory inference (dual-source and single-source), and the interaction between the two. The maximal converging model included random intercepts for participants and items, and random slopes for all fixed effects grouped by participant. We compared this model to a null model missing only the interaction of theoretical interest as a fixed effect. The models were significantly different (χ^2_1 = 7.9, $p < 0.01$), meaning the gap in difficulty for the memory-load task affected

dual-source and single-source illusory inferences differently, such that dual-source illusory inferences were more affected than single-source. Table 3 gives the coefficients of the full model.

Table 3

Coefficients of the full model

Coefficient	Estimate	Standard error
Intercept	1.6	0.3
Single-source inferences	−2.9	0.4
Easy difficulty	0.3	0.2
Interaction	−0.7	0.2

Two aspects of the pre-registered analysis we just reported were less than ideal. First, our mixed-effects logistic regressions did not converge with the largest possible set of random effects. This issue arises very frequently with these methods and does not *ipso facto* invalidate any conclusions, but it warrants flagging at any rate. Second, we noticed that the two standard information criteria, AIC and BIC, diverged with respect to our main analysis. Indeed, the full model, including the interaction term of interest, had a lower AIC than the null model (3777 vs. 3783, respectively), but it had a higher BIC, albeit only minimally so (3873 vs. 3872). Because lower values point to better models, the AIC but not the BIC supported the conclusion that dual-source illusory inferences are more affected by memory load than single-source inferences are. While this divergence did not appear in the analysis of large-scale pilot data we mentioned earlier (Picat, 2019), we decided to post-hoc redo our analysis with Bayesian mixed-effects models, as they are in general more likely to converge with large random-effects structures.

As with our pre-registered analysis, here we predicted participants' answers from the difficulty of the memory-load task, the class of illusory inferences, and the interaction between the two (our main factor of interest). The model included all random effects,

namely random intercepts for participants and items, as well as random slopes for all fixed effects grouped by participants and random slopes for memory-load grouped by items. As priors for all coefficients, we used an uninformative Student distribution with mean 0, standard deviation of 2.5 and 7 degrees of freedom. We compared this to a null model missing only the interaction. Using Leave-One-Out Cross Validation as a decision criterion, the model including the interaction comes out as better than the null model (difference in log pointwise predictive density being -5.3 with SD = 3.6).

In sum then, using *both* pre-registered frequentist methods *and* post-hoc Bayesian methods, we have evidence that dual-source illusory inferences are more affected by memory-load than single-source illusory inferences, confirming our main prediction. The effect we found was of a small size, as seen in Table 2 in the form of acceptance rates (drop of around 3 percentage points for dual-source inferences). This is of course quite subtle, but as we just argued there are good statistical reasons to think that this difference is indeed significant, and moreover a very small effect size was exactly what we expected given the very small effect sizes of the analogous manipulation in the study of scalar implicature (De Neys & Schaeken, 2007; Marty & Chemla, 2013).

Returning now to our pre-registered analysis for ancillary questions, we checked that each class of illusory inference led to fallacious conclusions in both difficulty conditions. We did this by comparing the answers to each of the two classes of illusory inferences with those to invalid controls, for each difficulty condition. For each of these four comparisons, we built a binomial linear mixed-effects model, including the type of inference (illusory inference or invalid control) as a fixed effect, random intercepts for participants and inference items, and a random slope for type of inference grouped by participant. Each model was compared to a corresponding null model missing the fixed effect of inference type. All model comparisons were significant, showing a main effect of inference type, such that more fallacious conclusions were drawn for illusory inferences than for invalid controls. Invalid controls establish a baseline for noisy mistakes, and

allowing the models to distinguish between invalid controls and illusory inferences improved them significantly. In other words, our results confirm that single-source and dual-source illusory inferences are fallacies in both the easy and the hard condition.

Table 4

Model comparisons testing if dual-source and single-source illusory inferences are fallacies in each difficulty condition

Class	Difficulty	Test statistic	*p*-value
Dual-source	Hard	$\chi^2_1 = 58.5$	$p < 0.0001$
	Easy	$\chi^2_1 = 69.3$	$p < 0.0001$
Single-source	Hard	$\chi^2_1 = 14.6$	$p < 0.001$
	Easy	$\chi^2_1 = 8.3$	$p < 0.01$

Table 5 gives the frequency of correct responses to valid and invalid control inferences, in each difficulty condition. Performance was excellent in both conditions.

Summing up, using a dual-task paradigm, we found that cognitive load affects dual-source and single-source illusory inferences differently, as predicted by our two-paths theory. The magnitude of the effect was similar to that found in studies applying the same kind of memory-load task to truth-value judgments about sentences that should normally trigger the interpretive processes of interest (De Neys & Schaeken, 2007; Marty & Chemla, 2013). The current results are in line with the work of Picat and Mascarenhas (2019) and they can be considered as a successful replication of the findings there.

As we suspected, the potential confound introduced by participants' possibly misparsing our universal target inferences as per (16b) (on page 25) does not obscure our results or our analysis: we found the effect even though one of our stimulus classes plausibly stacked the deck against us.

Table 5

Performance on control inferences

Type	Difficulty	Correct answers	Standard deviation	Standard error
Valid-control	Hard	96.2	10.0	0.9
	Easy	98.7	6.4	0.6
Invalid-control	Hard	87.8	22.3	1.9
	Easy	92.5	16.9	1.5

One *prima facie* plausible alternative account deserves discussion. Our manipulation affected single-source and dual-source illusory inferences differently, but looking at the frequencies in Tables 2 and 5 it seems that the manipulation might have affected every class of items, including controls, by pushing them toward chance levels. That is, across the board, the responses in the hard memory-load condition were closer to 50% than in the easy condition.

We question the plausibility of this alternative explanation of our results on both conceptual and empirical grounds. First, the dual-task paradigm we employed is in itself not novel, and has been used repeatedly in the psycholinguistics literature with results that corroborate certain theories and falsify others. It is highly unlikely that this was a product of chance and that the manipulation itself merely introduces noise. However, we made an unusual use of this methodology, by applying it not to single-sentence truth-value-judgment tasks, but to three-sentence inference tasks. It is possible that this modification of the experimental paradigm introduces significant noise. We cannot discard this alternative explanation purely on conceptual grounds.

We ran a post-hoc analysis to check the predictions of this alternative hypothesis. If noise explains our results, we would expect this noise to spread uniformly across all inference types. That is, the effect of the memory-load task should be the same for both

classes of illusory inferences and both types of controls. Accordingly, we checked whether dual-source illusory inferences were affected differently than valid controls by the memory-load task. We chose valid controls as a reference because their acceptance rate is closest to dual-source illusory inferences. We fitted participants' answers to valid controls and dual-source illusory inferences into a binomial linear mixed-effects model predicting answers from the type of inference, the difficulty of the memory-load task, and the interaction between the two. The maximal converging model included random intercepts for participants and items, and random slopes for inference type and difficulty grouped by participant. We compared this to a null model lacking the interaction coefficient as a fixed effect. On our two-paths theory, the model with the interaction should be better, since memory load is expected to affect dual-source and valid controls differently. On the alternative hypothesis we're considering, the models should be indistinguishable. If noise is to explain the difference between conditions, there is no reason for it to affect valid controls and dual-source inferences differently, given the fact that they have comparably high acceptance rates. We found a suggestive but only marginally significant difference between the models ($\chi^2_1 = 3.3, p = 0.07$).

In sum, we cannot at this point conclusively discard this alternative hypothesis. But we submit that a number of considerations provide arguments against it. First, our two-factor theory builds on independently motivated components co-opted from the two very distinct fields of linguistics and psychology. Its prior probability is relatively high. Second, the dual-task paradigm has repeatedly been proven effective in experiments similar to ours. Accordingly, the prior probability that noise is all that results from the manipulation in our experiment is low. Finally, the broader interaction study we ran post hoc was marginally significant, providing suggestive evidence that our manipulation treats dual-source illusions differently than valid controls, a distinction that a noise-based interpretation cannot make.

4 General discussion

When humans systematically draw fallacious conclusions from certain inference patterns, this could be for at least two broad classes of reasons. They could be engaging in a process of manipulating the available information that does not target validity in the sense of classical logic. This subsumes a diverse class of theories of reasoning. For example, humans might be making assessments of representativeness (Tversky & Kahneman, 1974), using non-classically valid derivation rules (Braine & O'Brien, 1998), or deploying a probabilistic notion of validity (Oaksford & Chater, 2007).

But something very different could also be happening. Humans might be reasoning with perfect respect for the normative standards of classical-logical validity, while working with non-obvious but rational and entirely predictable interpretations of the premises. Taking these interpretations into account, the purported mistake is no mistake at all, it only looked like one to the experimentalist and theorist who did not countenance interpretive processes with the appropriate degree of sophistication.

The existence of these two conceptual possibilities is a truism accepted by the entirety of the field since its earliest days. It is particularly well exemplified in the profusion of competing accounts of phenomena discovered by Daniel Kahneman and Amos Tversky, such as the conjunction fallacy.

Under certain circumstances, human reasoners will readily assign a higher probability to a conjunction than to one of its constitutive conjuncts, in apparent violation of the classical probability calculus. In the best-known variant of the experiment, participants were given a rich description of a character named Linda, including the information that she participated in social activist movements when she was in college. Participants were asked to judge which of two options was most probable: (A) Linda is a bank teller; (B) Linda is a bank teller and she is active in the feminist movement. About 85% of participants chose option (B).

While Tversky and Kahneman argued in favor of a reasoning-based account in

terms of typicality judgments for this phenomenon, they also considered the possibility of absolving interpretations. Structurally, if the isolated conjunct *a* is interpreted by contrast with the competing conjunction *a* AND *b*, then it is plausible for the isolated conjunct to be interpreted as *a* AND NOT *b*, in which case it may well have a lower probability than *a* AND *b*, without violating the axioms of classical probability theory (Tversky & Kahneman, 1983). These considerations marked only the very beginning of a vast literature on the interpretations of the conjuncts, the conjunction element itself (Hertwig & Gigerenzer, 1999), relative clauses (Mellers et al., 2001), and the intended notion of probability (Gigerenzer & Hoffrage, 1995), to name just a few.

The tension between reasoning proper and interpretation is universally acknowledged, but resolving it has been fraught with difficulty, as the theoretical quagmire that is the conjunction fallacy illustrates. We propose that two points in particular have played a key role preventing the field from successfully tackling the elusive line between reasoning and interpretation.

First, there is precious little work on multi-factor theories of reasoning phenomena incorporating both reasoning-based and interpretation-based factors. The literature on the conjunction fallacy again offers an instructive example. As we explained above, the original article by Tversky and Kahneman (1983) had already considered a reasonable absolving interpretation, building on classical theories of pragmatics. The authors controlled for this interpretation by changing the target sentences so as to block the relevant pragmatic processes. They still found conjunction errors, and concluded, with good justification, that this possible interpretation of their original stimuli did not altogether explain away the phenomenon. Yet they found a drop in conjunction mistakes in the order of 20 percentage points. And a careful discussion of this *partial* absolution obtained by controlling for interpretive processes is not to be found in the original article or in later research, to the best of our knowledge.

Second, there is not enough articulation between the two required expertises: the

psychology of reasoning and natural language semantics and pragmatics. Illusory inferences from disjunction have an interpretation-based account, but it requires the most modern approaches to pragmatics from the linguistics literature. If we had only considered the classical theories from philosophy of language and philosophical logic, as is standard in the reasoning literature, we would have missed the existence of an absolving interpretation for illusory inferences from disjunction.

Illusory inferences from disjunction and disjunction-like elements offer a particularly clear and well-behaved domain in which to address these two issues with current research into reasoning and interpretation. These illusions involve comparatively brief stimuli with no need for complex contexts, making their form simpler to control and their content easier to vary. They constitute classical validity tasks and thus connect straightforwardly to normative standards for rationality, an important feature when diagnosing *failures* of reasoning. And they involve linguistic elements that have been the object of sophisticated semantic and pragmatic analysis in recent years, namely disjunctions, indefinites, and epistemic modals. These elements form a unified class in some linguistic theories, but they are nevertheless diverse in form and meaning.

These conditions allowed us to spell out a two-factor account of these inferences that articulates appeals to general-purpose reasoning with pragmatic interpretation. The reasoning-based account, in terms of the erotetic theory of reasoning, predicts fallacies with disjunctions, indefinites, and certain modals. For disjunctions, pragmatics predicts the same inference-making behavior as the reasoning-based account, while seeing it not as a fallacy but as the result of sound reasoning operating on an enriched interpretation of the first premise. However, for indefinites and modals, pragmatics does not offer an absolving interpretation, and therefore does not predict the observed fallacious conclusion. Illusory inferences with disjunctions thus have two paths leading to the same inference-making behavior, while those with indefinites and modals have only one. Accordingly, disjunctive illusions have a very high acceptance rate, while indefinite and

modal illusions are appreciably less attractive.

We marshaled several arguments in defense of this view. First, we showed how articulating theories of interpretation from linguistics with theories of reasoning from psychology *made the prediction* that indeed these two kinds of inferences should exist and should display the degrees of attractiveness that they broadly display. Second, we observed that *existing experimental results* already pointed in this direction: Sablé-Meyer and Mascarenhas (2019) investigated a variant of our standard disjunctive inferences involving sets and supersets, for which an interpretation-based approach is unavailable. They found that their fallacies, while still quite attractive, were accepted far less often by human subjects, as the theoretical view we're pushing here would predict. Chung et al. (2022) translated the logical structure of the standard disjunctive illusory inference to a visual paradigm, and found that, while severely mitigated, the fallacy was still there, showing that removing language reduces but does not obviate the fallacious conclusion. These experimental results are very suggestive of the view we give here, but the variation they introduced changed the logical form of the problem a little too radically. The strongest possible demonstration of our theory would be found in an experimental manipulation that *blocked* the interpretive route to the fallacy while leaving the form of the reasoning problems *completely unchanged*. A tall order, we acknowledge, but one we believe we made significant strides toward.

The novel experiment in this article recruits a psycholinguistics method known to impede the pragmatic processes involved in the interpretation-based account. With it, we showed that dual-source illusory inferences (disjunctions) were more affected than single-source illusory inferences (indefinites and modals), with an effect size in the order of what had been found in the past in traditional psycholinguistic studies of pragmatics. This constitutes the first *completely direct* empirical evidence in favor of an interpretation-based account in dual-source illusory inferences from disjunction, and also the first empirical evidence that interpretive processes are involved to a different extent in

dual-source and single-source illusory inferences.

We did not offer experimental evidence that the lack of a pragmatic path *fully* explains the different acceptance rates of dual-source and single-source illusory inferences. The dual-task paradigm we employed serves as an effective diagnostic of pragmatic processes, but it is known not to fully block these pragmatic processes. A complete account will have to use other experimental paradigms that affect pragmatic processes with greater expected effect sizes. A promising candidate is priming. Bott and Chemla (2016) studied fine-grained mechanisms at play behind the interpretation of numerals and plurals, which have both received pragmatic accounts. They successfully use priming as a way to either increase or decrease the pragmatically strengthened interpretations of these elements, with effect-sizes in the order of 25 percentage points. This paradigm has certain disadvantages compared to our dual-task methodology, in particular the priming itself would rely on a task that is about language and reasoning, while our dual-task experiment manipulated pragmatic processes without using language, providing for a more elegant and less confounded design. On the other hand, experimental paradigms such as priming are likely to prove more fruitful when applied to broader classes of fallacies from the reasoning literature. It is hard for instance to imagine embedding the conjunction fallacy in a memory task, given the long descriptive paragraph about Linda.

Future research will tell which methodologies from psycholinguistics best lend themselves to being transposed to reasoning problems. We submit that this line of investigation holds great promise to illuminate the divide between reasoning proper and interpretive processes, if allied to a theoretical mindset that seeks articulated reasoning-based and interpretation-based accounts, and that looks both to psychology and to linguistics for theory building and experimental paradigms.

References

Alonso-Ovalle, L. (2006). *Disjunction in alternative semantics* (PhD diss.). UMass Amherst. https://semanticsarchive.net/Archive/TVkY2ZlM/alonso-ovalle2006.pdf

Anderson, J. R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences*, *14*(3), 471–517. https://doi.org/10.1017/S0140525X00070801

Bade, N., Picat, L., Chung, W., & Mascarenhas, S. (2022). Alternatives and attention in language and reasoning: A reply to Mascarenhas & Picat (2019). *Semantics & Pragmatics*. https://doi.org/10.3765/sp.15.2

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bott, L., & Chemla, E. (2016). Shared and distinct mechanisms in deriving linguistic enrichment. *Journal of Memory and Language*, *91*, 117–140. https://doi.org/10.1016/j.jml.2016.04.004

Braine, M. D. S., & O'Brien, D. (Eds.). (1998). *Mental logic*. Mahwah, NJ: Lawrence Erlbaum Associates.

Chierchia, G., Fox, D., & Spector, B. (2012). The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. In P. Portner, C. Maienborn, & K. von Heusinger (Eds.), *Semantics: An international handbook of natural language meaning*. Berlin: Mouton de Gruyter.

Chung, W., Bade, N., Blanc-Cuenca, S., & Mascarenhas, S. (2022). Question-answer dynamics in deductive fallacies without language. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*(44). https://escholarship.org/uc/item/9711612q

Ciardelli, I., Groenendijk, J., & Roelofsen, F. (2009). Attention! Might in inquisitive semantics. *Proceedings of the 19th Conference on Semantics and Linguistic Theory (SALT)*, 91–108. https://doi.org/10.3765/salt.v19i0.2520

Ciardelli, I., Groenendijk, J., & Roelofsen, F. (2019). *Inquisitive semantics*. Oxford University Press.

Crupi, V., Fitelson, B., & Tentori, K. (2008). Probability, confirmation, and the conjunction fallacy. *Thinking & Reasoning*, *14*(2), 182–199. https://doi.org/10.1080/13546780701643406

De Leeuw, J. R. (2015). Jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior research methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

De Neys, W., & Schaeken, W. (2007). When people are more logical under cognitive load: Dual task impact on scalar implicature. *Experimental Psychology*, *54*(2), 128–133. https://doi.org/10.1027/1618-3169.54.2.128

Gabry, J., Goodrich, B., & Lysy, M. (2020). Rstantools: Tools for developing r packages interfacing with stan [R package version version 2.1]. https://mc-stan.org/rstantools/

Gazdar, G. J. (1979). *Pragmatics: Implicature, presupposition, and logical form*. New York: Academic Press.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*(4), 684–704. https://doi.org/10.1037/0033-295X.102.4.684

Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2022). Rstanarm: Bayesian applied regression modeling via Stan. [R package version 2.21.3]. https://mc-stan.org/rstanarm/

Grice, P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics: Speech acts*. New York: Academic Press.

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

Hamblin, C. L. (1958). Questions. *Australasian Journal of Philosophy*, *36*(3), 159–168. https://doi.org/10.1080/00048405885200211

Hertwig, R., & Gigerenzer, G. (1999). The conjunction fallacy revisited: How intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making*, *12*, 275–305. https://doi.org/10.1002/(SICI)1099-0771(199912)12:4<275::AID-BDM323>3.0.CO;2-M

Hilton, D. J. (1995). The social context of reasoning: Conversational inference and rational judgment. *Psychological Bulletin*, *118*(2), 248–271. https://doi.org/10.1037/0033-2909.118.2.248

Horn, L. (1972). *On the semantic properties of the logical operators in English* (Doctoral dissertation). UCLA.

Jayaseelan, K. (2001). Questions and question-word incorporating quantifiers in Malayalam. *Syntax*, *4*(2), 63–93.

Jayaseelan, K. (2008). Question particles and disjunction [Unpublished manuscript: The English and Foreign Languages University (Hyderabad)]. http://ling.auf.net/lingBuzz/000644/current.pdf

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

Koralus, P., & Mascarenhas, S. (2013). The erotetic theory of reasoning: Bridges between formal semantics and the psychology of deductive inference. *Philosophical Perspectives*, *27*, 312–365. https://doi.org/10.1111/phpe.12029

Koralus, P., & Mascarenhas, S. (2018). Illusory inferences in a question-based theory of reasoning. In K. Turner & L. Horn (Eds.), *Pragmatics, truth, and underspecification: Towards an atlas of meaning* (pp. 300–322). Leiden: Brill. https://doi.org/10.1163/9789004365445_011

Kratzer, A., & Shimoyama, J. (2017). Indeterminate pronouns: The view from Japanese. In C. Lee, F. Kiefer, & M. Krifka (Eds.), *Contrastiveness in information structure, alternatives and scalar implicatures* (pp. 123–143). Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-10106-4_7 (Original work published 2002)

Marty, P. P., & Chemla, E. (2013). Scalar implicatures: Working memory and a comparison with *only*. *Frontiers in Psychology*, *4*(403). https://doi.org/10.3389/fpsyg.2013.00403

Marty, P. P., Chemla, E., & Spector, B. (2013). Interpreting numerals and scalar items under memory load. *Lingua*, *133*, 152–163. https://doi.org/10.1016/j.lingua.2013.03.006

Mascarenhas, S. (2009). *Inquisitive semantics and logic* (Master's thesis). ILLC. https://eprints.illc.uva.nl/id/eprint/825/1/MoL-2009-18.text.pdf

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

Mascarenhas, S. (2014). *Formal semantics and the psychology of reasoning: Building new bridges and investigating interactions* (Doctoral dissertation). New York University. https://ling.auf.net/lingbuzz/002213

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

Mascarenhas, S., & Picat, L. (2019). *Might* as a generator of alternatives: The view from reasoning. *Proceedings of SALT 29, UCLA*, 549–561. https://doi.org/10.3765/salt.v29i0.4635

Mastropasqua, T., Crupi, V., & Tentori, K. (2010). Broadening the study of inductive reasoning: Confirmation judgments with uncertain evidence. *Memory and Cognition*, *38*(7), 941–950. https://doi.org/10.3758/MC.38.7.941

Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science*, *12*(4), 269–275. https://doi.org/10.1111/1467-9280.00350

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

Over, D. E. (2009). New paradigm psychology of reasoning. *Thinking & Reasoning*, *15*(4), 431–438. https://doi.org/10.1080/13546780903266188

Picat, L. (2019). *Inferences with disjunction, interpretation or reasoning?* (Master's thesis). Ecole Normale Supérieure. http://web-risc.ens.fr/~lpicat/website/picat-m2-thesis-pre-print.pdf

Picat, L., & Mascarenhas, S. (2019). Reasoning with disjunctions under cognitive load [Talk given at *Brain, Language and Learning* 2019, University of Siena].

Picat, L., & Mascarenhas, S. (2020a). Reasoning with linguistic alternatives under cognitive load [Preregistration document available on OSF]. osf.io/9dpbw

Picat, L., & Mascarenhas, S. (2020b). Reasoning with linguistic alternatives under cognitive load [OSF repository with all materials and code used to run this experiment]. osf.io/976w8/files/

R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

RStudio Team. (2016). *Rstudio: Integrated development environment for r*. RStudio, Inc. Boston, MA. http://www.rstudio.com/

Sablé-Meyer, M., & Mascarenhas, S. (2019). Assessing the role of matching bias in reasoning with disjunctions [Poster presented at the 41st Annual Meeting of the Cognitive Science Society]. http://web-risc.ens.fr/~smascarenhas/docs/sable-meyer-mascarenhas19_matching_bias_disjunctions.pdf

Sablé-Meyer, M., & Mascarenhas, S. (2021). Indirect illusory inferences from disjunction: A new bridge between deductive inference and representativeness. *Review of Philosophy and Psychology*, *12*(2). https://doi.org/10.1007/s13164-021-00543-8

Sauerland, U. (2004). Scalar implicatures in complex sentences. *Linguistics and Philosophy*, *27*, 367–391. https://doi.org/10.1023/B:LING.0000023378.71748.db

Spector, B. (2007). Scalar implicatures: Exhaustivity and Gricean reasoning. In M. Aloni, P. Dekker, & A. Butler (Eds.), *Questions in dynamic semantics*. Elsevier.

Stan Development Team. (2023). RStan: The R interface to Stan [R package version 2.21.8]. https://mc-stan.org/

Stenning, K., & van Lambalgen, M. (2008). Interpretation, representation, and deductive reasoning. In J. Adler & L. Rips (Eds.), *Reasoning: Studies in human inference and its foundations* (pp. 223–249). Cambridge: Cambridge University Press.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293–315. https://doi.org/10.1037/0033-295X.90.4.293

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., & Gelman, A. (2022). Loo: Efficient leave-one-out cross-validation and waic for bayesian models [R package version 2.5.1]. https://mc-stan.org/loo/

Walsh, C., & Johnson-Laird, P. N. (2004). Co-reference and reasoning. *Memory and Cognition*, *32*, 96–106. https://doi.org/10.3758/BF03195823

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., . . . Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686