Illusory inferences in a question-based theory of reasoning

Philipp Koralus[a] and Salvador Mascarenhas[a]

[a]Laboratory for the Philosophy and Psychology of Rationality and Decision (LPPRD)
Faculty of Philosophy
University of Oxford
Radcliffe Observatory Quarter
Woodstock Road
Oxford, OX2 6GG
UK
Corresponding author: philipp.koralus@stcatz.ox.ac.uk

Abstract

The erotetic, or question-based, theory of reasoning has been developed from an account of utterance interpretation but moves beyond interpretation to a general account of human reasoning with multiple premises that covers both ideal valid reasoning as well as fallacious reasoning. The core idea of the erotetic theory of reasoning is that reasoners treat premises as questions and maximally strong answers, which can be made mathematically precise using tools from philosophy and linguistics. We present three experiments showing new fallacious inferences predicted by the erotetic theory, some of which are so frequently endorsed to be akin to cognitive illusions. The results suggest that similar phenomena discussed in the literature are more general and systematic than had been thought.

1 Introduction

The capacity for reasoning is central to modern human endeavors. Philosophers and linguists have often focused on idealized forms of reasoning as part of an account of utterance interpretation. Systematic consideration of reasoning as a cognitive capacity has largely been left to psychologists. The most widely discussed approaches to our reasoning capacity are mental logic (Rips, 1994), Bayesianism (Oaksford and Chater, 2007), and mental model theory (Johnson-Laird, 1983). Each of these approaches has important advantages. Mental logic approaches incorporate the insight that a fully satisfactory theory of reasoning should be formally precise in the sense that it is possible to calculate predictions of the theory for an unbounded number of reasoning problems from a clearly specified set of axioms. Bayesian approaches have the advantage that they seek to account for our reasoning performance in ways that would flow from a simple core idea of what our reasoning capacity is aiming at, e.g. rationally updating a probability distribution in light of evidence. Mental model theory has the advantage of making central the nature of the representations we build as we interpret premises and how those representations can be less detailed than the premise statements in terms of their information content. We have argued for a new theory of reasoning that we believe unites those advantages, the erotetic theory of reasoning (ETR) (Koralus and Mascarenhas, 2013), taking cues from a related account of utterance interpretation (Koralus, 2012) but moving beyond interpretation to reasoning proper. The core idea behind ETR is the notion that we reason by raising questions and seeking to answer them as directly as possible. This intuitive notion has been given a formally rigorous description (Koralus and Mascarenhas, 2013) elsewhere. In this paper, we focus on presenting three new experiments on reasoning documenting novel systematic fallacies predicted by the erotetic theory. We thus broadly follow a vision according to which tools primarily used in the study of language can yield insights about cognition more broadly (Atlas, 2005).

Inferences from disjunctive statements involving ‘or’ are among the simplest non-trivial cases of reasoning. For example, if we accept that either there is an apple on the table or else an orange, and we further accept that there isn’t an orange on the table, we may straightforwardly conclude that there is an apple. Now, we are subject to systematic fallacies of reasoning with premises that are similarly simple. Walsh and Johnson-Laird (2004) presented participants in an experiment with the following problem:

(1) **(P1) 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.**
(P2) Jane is kneeling by the fire.
Does it follow that she is looking at the TV?

Remarkably, only about 10% of participants gave correct answers to problems of this form, prompting the authors to call them “illusory inferences.”[1] Most participants say that it follows from the two premises that Jane is looking at the TV. However, on reflection, we can see that this is a fallacy. The truth of the premises is compatible with a situation in which Jane is kneeling by the fire but *not* looking at the TV, while Mark is standing by the window and is peering into the garden. The general pattern of these types of fallacies is

[1] A correct inference from the premises in (1) would be, for example, any restatement of the premises, as well as a statement that “nothing follows.”

not readily explained by recourse to pragmatic interpretive factors. For example, the conclusion that Jane is looking at the TV still does not follow under an exclusive-or interpretation of ‘or’. We will say more about more sophisticated pragmatic accounts toward the end of the paper after discussing our experiments and looking at the relevant patterns in greater generality.

To date, the only systematic accounts of illusory inferences of this kind are offered by mental model theory and by the erotetic theory of reasoning. Both theories account not only for such inferences with disjunctions and conjunctions as in (1) (mental models theory: Walsh and Johnson-Laird, 2004; erotetic theory: Koralus and Mascarenhas, 2013), but also for illusory inferences with conditionals (MMT: Johnson-Laird and Savary, 1999; ETR: Koralus and Mascarenhas, 2013) and with quantifiers (MMT: Khemlani and Johnson-Laird, 2012; ETR: Mascarenhas and Koralus, 2015). The mental-model based explanation Walsh and Johnson-Laird propose for this illusory inference is based on the idea that we build mental models of only some of the alternative possibilities compatible with the first premise, notably of the alternatives in which Jane is kneeling and looking, the alternative in which Mark is standing and peering, and the alternative which combines both of the former. Walsh and Johnson-Laird suggest that when we encounter the second premise we “match” that premise to the alternatives from the first premise that are partly co-referential with it. This match is then treated as definitively establishing that one of these alternatives holds. In this case, all envisaged alternatives that “match” Jane kneeling by the fire also include that Jane is looking at the TV, yielding the illusory inference.

ETR proposes a rather different explanation of illusory inferences. The core idea is that a disjunctive premise statement raises the question of which of the disjuncts is the case. Effectively, (P1) is interpreted as akin to the question, “am I in a kneeling and looking situation or in a standing and peering situation?” (P2) is then taken as akin to “you are in a kneeling situation!”, *interpreted as a maximally strong answer*. Taking P2 as a maximally strong answer to P1 would lead us to conclude that we are in a kneeling and looking situation. This question/answer-based explanation of the illusory inference pattern does not require co-reference between expressions in the premise statements. Moreover, it predicts an order effect and other cases of “illusory” inferences, which we address in experiments reported further below. Our results show that the pattern of “illusory” inferences is much more general and systematic than has previously been reported. We argue that existing theories do not account for this data and use this to motivate the notion of reasoning as question-answering. We end the paper with a description of how this idea yields a new general account of propositional reasoning that can make sense of both fallacies and of the possibility of valid reasoning.

2 Experiment 1

2.1 Illusory inferences from disjunctions without co-reference

Our first concern was to establish whether the illusory inference pattern is in fact due to co-reference as Walsh and Johnson-Laird (2004) suggest, or whether it applies more broadly to premise pairs of the form "(P1) A and B or B and C. (P2) A." To do this, we relied on indefinite expressions. Successive uses of indefinites like 'a' and 'an' are not typically interpreted as referring to the same individual (Heim, 1982), so their use would not provide for interpretations on which successive premises are linked by co-reference, unlike expressions like proper names and indexical expression like 'it', which were used by Walsh and Johnson-Laird (2004).

In experiment 1, we examined whether illusory inferences from disjunction are driven by processes that are specific to building mental models with co-reference across premises. We examined four illusory inference problems and four control problems that were not hypothesized to yield illusory inferences. Both types of problems involved two premises, where the first premise consisted of a disjunction and the second premise consisted of an atomic proposition or a negated atomic proposition. The target and control problems were variants of the following two examples:

(2) **Sample Target Problem**.

There is an ace and a queen, or else a king and a ten.

There is a king.

What if anything follows?

(3) **Sample Control Problem**.

There is an ace and a king, or else a queen and a jack.

There isn't an ace.

What if anything follows?

We predicted that, in the target problem, participants should systematically draw the illusory inference that there is a ten in the hand, and so on in similar problems. We predicted that the incidence of these mistakes in target problems should be far greater than the incidence of invalid inferences in the control problems.

2.2 Method

Participants and Design. 241 members of the mTurk worker community (average age 33 years, $\sigma = 10.2$, 94 female, 146 male) carried out eight reasoning problems, including the same four target and four control problems for each subject. Subjects were randomly assigned to one of two conditions. In condition 1 (126 subjects), subjects saw problems as in (2) and (3) above, in condition 2 (115 subjects), subjects saw the same materials with premises in reversed order. In this section we reports the results of condition 1, the canonical order version. Subjects served as their own controls. The order of presentation was randomized for each subject. Both target and control problems each consisted of a disjunctive premise with two disjuncts followed by a premise of either atomic nor negated atomic propositional form. In each problem, the premise statements were followed by the question, "what if anything follows?" and a text box to record responses.

Procedure. The experiment was carried out over the internet using Qualtrics and participants were anonymously recruited and paid through the Amazon mTurk website. Each participant was rewarded with USD 0.25 for their participation. Participants were invited to engage in a study of reasoning in which they had to say what they can conclude from a set of statements. They were asked not to make notes or use search engines while performing the task. Before the target and control questions were presented, each participant was shown two worked-out sample reasoning problems of an unrelated kind using conditionals. All statements were explained to be about a large hand of cards. Participants typed their responses into text boxes under the premise statements. They were told that the experiment would last approximately 5 minutes but were given as much time as they needed, up to 10 minutes. On average, they took just under five minutes to complete the experiment.

2.3 Results

The participants' written responses were coded as follows into binary categories. For target problems, we coded a response as a "1" in the category of illusory inference (ILL) if and only if at least one illusory inference proposition was a conjunct in the written answer and no other invalid inferences were present. For control problems, we coded a response as invalid (INV) if and only if at least one invalid inference of any sort was present, excepting responses like "nothing follows," "no," and the like. In all cases, we made allowance for the fact that some participants may interpret 'or' as inclusive and some as exclusive.

Type	Pattern	ILL	INV
Target	aq ∨ kx ‖ k	116(92%)	-
Target	j2 ∨ a8 ‖ j	115(91%)	-
Target	q8 ∨ 2a ‖ 2	116(92%)	-
Target	xk ∨ qa ‖ x	114(90%)	-
Control	ak ∨ qj ‖ ¬a	-	14(11%)
Control	j8 ∨ ax ‖ ¬j	-	19(15%)
Control	28 ∨ kj ‖ ¬k	-	22(17%)
Control	8k ∨ qa ‖ ¬q	-	22(17%)

Table 1: Results of experiment 1, condition 1. *"a", "j", "k", "q", "2", "8", and "x" stand for the cards ace, jack, king, queen, two, eight, and ten, respectively.*

Two research assistants, unaware of the study's hypotheses, coded the free-form responses into the above categories, agreeing on 99.2% of the data points. The few discrepancies were resolved by a third coder. 98% of participants made one or more illusory inferences, while 22% made one or more invalid inferences in control problems. We rejected the null hypothesis that illusory inferences were as frequent as invalid inferences in control problems (Wilcoxon Matched Pairs Test, $V = 83,400$, $p < 0.0001$). We summarize the data for the different target and control problems in Table 1.

2.4 Discussion

Our results show that the pattern of illusory inference is not due to co-reference across premises. The explanation offered by ETR does not depend co-reference. ETR holds instead that the illusory inference results from treating the first premise of our sample target in (2) as a question, e.g. "am I in an ace and queen situation or in a king and ten situation?," and treating the second premise as akin to "you're in a king situation," *interpreted as a maximally strong answer*, yielding the fallacious conclusion that there is a king and a ten.

2.5 Order effect for illusory inference from disjunction

One could take the Walsh and Johnson-Laird (2004) notion of "matching" to extend to atomic propositions like "there is an ace," regardless of co-reference (as opposed to more abstract models). However, it is worth

Type	Pattern	ILL
Target	k ‖ aq ∨ kx	90(78%)
Target	j ‖ j2 ∨ a8	95(82%)
Target	2 ‖ q8 ∨ 2a	90(78%)
Target	x ‖ xk ∨ qa	88(77%)

Table 2: Results of experiment 1, condition 2. *"a", "j", "k", "q", "2", "8", and "x" stand for the cards ace, jack, king, queen, two, eight, and ten, respectively.*

noting that matching does not intrinsically invoke order. Matching curtains to a sofa and matching a sofa to curtains should yield the same outcome if there is only one possible matching sofa-curtain pair (e.g., in the problems we considered, there is only one possible match across the two premises). By contrast, we cannot treat something as an answer without having a question first. Thus, the explanation of the illusory inference offered by ETR, unlike the matching-based account, immediately predicts an order effect. If the premises in the illusory inference problems are reversed, the illusory inference should be mitigated. Condition 2 of experiment 1, where subjects saw the same materials as in condition 1 with the order of the premises reversed, allowed us to test this hypothesis.

2.6 Results

Participants' responses were coded into the same categories as condition 1 described above. The two initial coders agreed on 97.7% of data points, and disagreements were resolved by a third coder. As predicted, fewer illusory inferences were made in the reversed condition compared to the canonical-order condition in experiment 1. The number of illusory inferences dropped by approximately 10% on average when the premises were reversed. We rejected the null hypothesis that the number of illusory inferences was the same for the two premise orders (Mann-Whitney, $W = 130,474$, $p < 0.0001$). There was no significant effect of premise order in our controls (disjunctive syllogism, Mann-Whitney, $W = 117,502$, $p > 0.5$). The results for the target cases are summarized in Table 2.

2.7 Discussion

A drop in acceptance rates for the reversed illusory inferences in the order of 10% does not entail of course that illusory inferences disappear when premises are reversed. But clearly the reversal has a significant ameliorating effect that was not found in our controls. These results are consistent with the prediction of ETR that reversing the order of premises should mitigate the illusory inference pattern. They cast doubt on explanations of these kinds of inference patterns involving notions like "matching" that are not relevantly asymmetric.

Now, mental model theory also provides a more general procedure for conjoining mental models (Johnson-Laird, 2008; Khemlani and Johnson-Laird, 2009). If we apply this to the problems in this experiment, the first premise would generate a single mental model, which would then have to be combined with the set of mental models generated by the second premise according to the process for "conjoining" mental models (*ibid.*). However, this process is also designed to make us jump to conclusions in reasoning. As we conjoin the mental model for "there is a king" with the mental models for "there is an ace and a queen, or else a king and a ten," the procedure, as defined by Johnson-Laird and Khemlani, rules out all of those mental models in which we do not have an ace and a queen. Thus, the mental model conjoin procedure applied to the case at hand would also yield the illusory inference in the reversed case. Moreover, the mental model conjoin procedure would *not* predict an illusory inference in the canonical order for which the illusory inference is in fact the strongest (see experiment 1). In sum, we do not see a non-ad hoc way to account for the observed order effect with the classical mental model theory.

3 Experiment 2

Illusory inferences from two disjunctions

In experiment 2, we further investigated the nature of illusory inferences from disjunctions. If it is crucial for obtaining an illusory inference that we match a *categorical* premise to one of the alternative models induced by a disjunctive premise, we would not expect illusory inferences from *two* disjunctive premises. However, if we take the view that what explains illusory inferences is a process of treating successive premises and questions and maximally strong answer to them, as formally defined in Koralus and Mascarenhas (2013), there is no such restriction. We can clearly treat disjunctions as answers to questions (e.g. Q: Where's Mary? A: Either at home or at work.).

We hypothesized that a significant number of participants would draw an illusory conclusion from premises like the following.

Type	Pattern	ILL	INV
Target	aq ∨ kx ‖ k ∨ k5	62(50%)	-
Target	j2 ∨ a8 ‖ j ∨ j7	62(50%)	-
Target	q8 ∨ 2a ‖ 2 ∨ 26	63(51%)	-
Target	xk ∨ qa ‖ x ∨ x9	61(49%)	-

Table 3: Results of experiment 2. *"a", "j", "k", "q", "2", "8", and "x" stand for the cards ace, jack, king, queen, two, eight, and ten, respectively.*

(4) **Sample Target Problem**.

There is a queen and an eight, or else a two and an ace.

There is a two, or a two and a six.

What if anything follows?

In the above example, the hypothesized illusory inference would be to say that it follows that there is an ace. The procedures were the same as in experiment 1, using the same control problems. A new set of 124 participants was recruited (average age 36, $\sigma = 10.8$, 46 female, 75 male).

3.1 Results

Free form responses were coded by two blind coders as in the previous experiment, who agreed on 95.4% of the data. Disagreements were resolved by a third coder. 60% of participants made an illusory inference in at least one of the four target problems, while 23% made an invalid inference in at least one control problem. We rejected the null hypothesis that the frequency of illusory inferences on target problems was the same as that of invalid inferences in control problems (Wilcoxon Matched Pairs Test, $V = 33,841.5$, $p < 0.0001$). The results for target problems are summarized in Table 3 (control problem results omitted since they were nearly identical to those in experiment 1).

3.2 Discussion

We found that 60% of participants made at least one illusory inference from two disjunctive premises. This shows that it is not a requirement for illusory inferences from disjunctions that one have a categorical premise to match to an alternative possibility provided by the disjunctive premise. The mental model conjoin procedure described in Khemlani and Johnson-Laird (2009) does not generate this illusory inference either, so it looks like none of the mechanisms in classical mental model theory for generating illusory inferences capture these data. Curiously, the mental model conjoin procedure *would* generate the illusory inference if the order of the premises was reversed.

ETR straightforwardly predicts the observed inference pattern, as can be computed from the formal system in Koralus and Mascarenhas (2013). Intuitively, disjunctions that have categorical entailments can serve as answers. In the case of the example in (4), "there is a two or a two and a six" answers the question of whether there is a two.

So far, we have only seen illusory inferences from disjunctions that were driven by hastily eliminating alternatives obtained from a first premise that in some sense did not overlap with the information in a further premise. Experiment 4 addressed whether this notion needs greater generality.

4 Experiment 3 Illusory inferences from triple disjunctions

In experiment 3, we examined a further novel illusory inference pattern. The question motivating this experiment is whether we make illusory inferences because we detect *overlap* between an alternative in a first premise with what its established by a second premise, or whether we look for alternatives in a first premise that has the most in common with what is established by further premises. The thought is that if we treat further premises as *maximally strong* answers, we might expect that only alternatives in the question that have equally much in common with the answer survive. This is an appreciably more powerful notion that would correspondingly yield fallacies not predicted by mere "overlap." Both target and control problems involved three premises, where the first premise consisted of a triple disjunction and the second and third premises consisted of conjunctions or negated atomic statements. The control problems had valid non-disjunctive conclusions that could be reached by two disjunctive-syllogism inferences. The target problems did not have valid non-disjunctive conclusions besides the trivial restatement of the second or third premises. The target and control problems were variants of the following two examples:

(5) **Sample Target Problem**.

There is an ace and a jack and a queen, or else there is an eight and a ten and a two, or else there is an ace.

There is an ace and a jack, and there is an eight and a ten.

There is not a queen.

What if anything follows?

(6) **Sample Control Problem**.

There is an ace and a king and a queen, or else there is an ace and a jack and a ten, or else there is an eight.

There isn't queen.

There isn't a ten.

What if anything follows?

For the target problem, we predicted that participants would systematically draw the illusory inference that there is a two in the hand. We predicted furthermore that the incidence of these mistakes in target problems should be far greater than the incidence of invalid inferences in the control problems.

4.1 Method

The protocol followed was the same as that in experiment 1, with a new set of participants (121, average age 33 years, $\sigma = 10.5$, 70 Female, 51 Male).

4.2 Results

Free form responses were coded by two blind coders as in the previous experiment, who agreed on 98.2% of the data. Disagreements were resolved by a third coder. 79% of participants made an illusory inference in at least one of the four target problems, while 25% made an invalid inference in a control problem. We rejected the null hypothesis that illusory inferences were as frequent as invalid inferences in control problems (Wilcoxon Matched Pairs Test, $V = 39,243$, $p < .0001$). We summarize the data for the different target and control problems in Table 4 on the next page.

4.3 Discussion

The illusory inference seen in experiment 3 shows that any procedure that is entirely based on eliminating alternatives in disjunctive premises depending on whether they fail to overlap with a categorical premise

Type	Pattern	ILL	INV
Target	ajq ∨ 8X2 ∨ a ‖ aj8X ‖ ¬q	67(55%)	-
Target	qkX ∨ ja8 ∨ q ‖ qkja ‖ ¬X	70(58%)	-
Target	kjX ∨ qa2 ∨ k ‖ kjqa ‖ ¬2	74(61%)	-
Target	j8q ∨ 2ka ∨ j ‖ j82k ‖ ¬a	65(54%)	-
Control	akq ∨ ajX ∨ 8 ‖ ¬q ‖ ¬X	-	20(17%)
Control	jk8 ∨ qaX ∨ 2 ‖ ¬8 ‖ ¬X	-	21(17%)
Control	qjX ∨ ka8 ∨ 2 ‖ ¬8 ‖ ¬2	-	19(16%)
Control	X82 ∨ jkq ∨ a ‖ ¬2 ‖ ¬a	-	16(13%)

Table 4: Results of experiment 3. *"a", "j", "k", "q", "2", "8", and "x" stand for the cards ace, jack, king, queen, two, eight, and ten, respectively.*

fails to account for illusory inferences in their full generality. *Every* alternative in the first premise overlaps with something in the second premise, so just checking *whether* there is overlap cannot give us the fallacy. The mental model conjoin procedure fails to capture the fallacious inference. However, if we generalize the central idea of the erotetic theory of reasoning according to which we treat successive premises as questions and strongest-possible answers, we get the right result. We propose that as we consider the three disjuncts of the first premise, we effectively are asking, "are we in an ace & jack & queen situation, an 8 & 10 & 2 situation, or an ace situation?" Treating the second premise as a maximally strong answer, we take those alternatives to be ruled out that have the *least* in common with a situation in which there is an ace & jack & 8 & 10. This narrows the three alternatives down to two. Finally, the third premise narrows what remains down to one alternative, yielding the fallacious conclusion that there is a two.

With the last experiment in place, we can return to the issue of pragmatic alternative explanations. We noted earlier that a straightforward analysis in terms of exclusive interpretations of disjunction will not suffice to explain away the fallacies we report as artifacts of interpretation. But modern approaches to formal pragmatics go well beyond what can be captured with exclusive 'or', and could in principle do the required job for the pattern in experiment 1. In fact, Mascarenhas (2014) proved that a relatively wide range of contemporary pragmatic theories can in fact account for Walsh and Johnson-Laird's (2004) original illusory inferences as matter of interpretation. As it turns out, most theories of formal pragmatics predict a much stronger interpretation for the first premise of classical illusory inferences from disjunction than simple exclusive disjunction. (7-a) shows the interpretation of the first premise, before pragmatic strengthening, and (7-b) the interpretation after strengthening. This strengthening is expected under all major theories, such as Sauerland (2004) or Spector (2007).

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

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

Under the interpretation in (7-b) and together with the second premise a, the (no longer) "fallacious" conclusion b follows validly by disjunctive syllogism and conjunction elimination. This story works as well for the novel pattern we introduced in experiment 2 with a disjunctive second premise. Second premises in experiment 2 were of the shape $a \vee (a \wedge b)$, which, being classically equivalent to a, will combine with the strengthened meaning of the first premise exactly as in the case just discussed.

Scalar implicatures may contribute to the phenomena discussed here in some cases, but they do not yield a satisfactory general account. The pragmatic account has nothing to say about our data in experiment 3, with three premises. For concreteness, we schematize the stimulus in (5) from experiment 3:

(8) P1: $(a \wedge j \wedge q) \vee (8 \wedge X \wedge 2) \vee a$

P2: $a \wedge j \wedge 8 \wedge X$

P3: $\neg q$

Concl: 2

Premises 2 and 3 do not involve low scalar items, so we can assume that their strengthenings are trivial. Premise 1 however is of some interest, as in the simpler cases discussed above. In principle, it should be interpreted in a strongly exclusive fashion, where negations of the atomic propositions themselves will figure into the strengthening meaning, as follows.

(9) $(a \wedge j \wedge q \wedge \neg 8 \wedge \neg X \wedge \neg 2) \vee (8 \wedge X \wedge 2 \wedge \neg a \neg j \wedge \neg q) \vee (a \wedge \neg j \wedge \neg q \wedge \neg 8 \wedge \neg X \wedge \neg 2)$

Now, if premise 1 is interpreted as in (9), the set of the premises would be inconsistent, for each disjunct of (9) contradicts some atom entailed by premise 2 of (8). This is clearly not an avenue for analysis, but perhaps all is not yet lost. Scalar implicatures cannot contradict asserted material, as evidenced by the fact that, if (10-a) is part of the common ground, an assertion of (10-b) will not constitute a contradiction of what is established, rather it will lack the scalar implicature it typically carries.

(10) a. John graded all of the scripts.

b. John graded some of the scripts.

Plausibly then, reasoners strengthen the first premise as in (9), but when confronted with contradictory information *asserted* by the second premise, they backtrack and excise from the strengthened interpretation of the first premise all negations of newly asserted atomic propositions. The set of premises would be strengthened as in (11), where as before premises 2 and 3 are assumed to be unchanged by pragmatic processes.

(11) P1: $(a \wedge j \wedge q \wedge \neg 2) \vee (8 \wedge X \wedge 2 \wedge \neg q) \vee (a \wedge \neg q \wedge \neg 2)$

P2: $a \wedge j \wedge 8 \wedge X$

P3: $\neg q$

Concl: 2

The conjunction of the premises of (11) is now consistent, but notice that the observed conclusion that there is a two in the hand does not follow. We conclude that our novel data in experiment 3 are not amenable to a pragmatic account.

The problems for a general pragmatic account of these kinds of inferences are not limited to the data in experiment 3. Mascarenhas (2014) showed that scalar implicature is in principle incapable of accounting for versions of illusory inferences with indefinite quantifiers doing the job of disjunction. We refer the reader to Mascarenhas (2014) and to Mascarenhas and Koralus (2015) for an exposition of these illusory inferences with quantifiers and the issues with scalar implicature accounts.

5 General discussion — the erotetic theory of reasoning

We have presented several novel patterns of illusory inference and we have argued that existing accounts fail to capture these patterns. Now, we want to maintain the insight from Johnson-Laird and his collaborators that reasoning proceeds by building mental models of premise statements. In that sense, we argue against a very specific "mental model theory" that we discussed above while wholly embracing the idea that reasoning is based on mental models. However, we propose a novel view of what these models contain and of how mental models are updated when successive premise statements are taken into account.

On the erotetic theory of reasoning, mental models are updated with the aim of answering the questions they represent. In Koralus & Mascarenhas (2013) we provide a formally complete presentation of the theory for the case of propositional reasoning. This theory is compatible with recent work in linguistic semantics within the frameworks of alternative semantics (Kratzer and Shimoyama, 2002) and inquisitive semantics (Groenendijk, 2008; Mascarenhas, 2009).

There, we also show how the erotetic theory accounts for the rich collection of exiting data on reasoning with propositional connectives. Some of the most relevant data points are summarized in Table 5 at the end of this paper.

The central idea of the erotetic theory of reasoning is that reasoning proceeds by updating an integrated mental representation of alternative possibilities in light of successive premise statements. By default, this process of updating proceeds by treating successive premises as questions and maximally strong answers to them (Part I of the erotetic principle below). Statements are interpreted relative to a question that a hearer or reasoner seeks to answer in a way that goes beyond the narrow propositional contribution of the answer (Koralus, 2012). The core of the theory is summarized in the erotetic principle in (12).

(12) **The erotetic principle**

Part I — Our natural capacity for reasoning proceeds by treating successive premises as questions and maximally strong answers to them.

Part II — Systematically asking a certain type of question as we interpret each new premise allows us to reason in a classically valid way.

Part II of the erotetic principle concerns itself with answering what one might call the problem of success for human reasoning. Humans are not irretrievably lost to the non-normative conclusions brought about by their tendency to try to find immediate strong answers. This fact needs to be explained alongside whatever failures of reasoning we exhibit. On the erotetic theory, questions in fact also play a crucial role in leading us to normatively correct reasoning. If reasoners raise enough questions that would force them to consider alternatives that they would otherwise neglect, then their reasoning is guaranteed to be sound. In particular, if reasoners ask polar questions (i.e. yes-no questions) about each atomic proposition that occurs in the question under consideration *before* updating with the putative answer supplied by a later premise, it can be shown that their reasoning will be classically sound in the technical sense. We prove this result as a theorem in Koralus and Mascarenhas (2013). We will end this section with an overview of the components of the theory of propositional reasoning laid out fully in Koralus and Mascarenhas (2013).

5.1 Key components of the theory

5.1.1 A theory of mental representations

The first step is to specify what contribution individual premise statements make. We adopt the view that ‘or’ raises the question of which of the disjuncts is the case. This is in line with much recent work in linguistic semantics within the frameworks of alternative semantics (Kratzer and Shimoyama, 2002) and inquisitive semantics (Groenendijk, 2008; Mascarenhas, 2009). The key insight of these approches to the meaning of disjunction is that there is much to be gained from taking ‘or’ and questions to denote the same kinds of mathematical objects. This interpretive move is independently motivated by the fact that in many natural languages the morphemes for disjunction and question formation are etymologically related (Mascarenhas, 2009).[2] Following standard approaches, we model questions as sets of alternative answers. For example, for a premise statement like “There is an ace and a queen or a king and a jack”, we obtain the set $\{a\&q, k\&j\}$. Why do we take it that the question we get from a disjunction includes these alternatives but not others? We propose that the alternatives we represent are the *exact verifiers* or “truth-makers” of the disjunction (**?**). What this means is that we only include as alternatives things that exactly make the statement true and no more. For example, $a\&q\&j$ would make “There is an ace and a queen or a king and a jack” true. However, it wouldn’t do so exactly. Clearly, we could drop j from $a\&q\&j$ and still make the statement true. As we will discuss in the section on “inquiring,” exact verifier representations do not in fact involve throwing away information. However, they create the possibility of reasoners ignoring certain possibilities compatible with their premises by simply not representing them as explicit alternatives.

Moving on, for a simple premise like “there is an ace,” we obtain a singleton $\{a\}$, after all, you exactly need an ace to make “there is an ace” true. A fully systematic specification of how interpretations are obtained from premises in propositional reasoning problems can be found in Koralus and Mascarenhas (2013).

[2] Sentences with disjunctions are not the only superficially declarative sentences that we take, with the lingustics literature cited, to share the crucial property with questions. In particular, indefinite expressions should also give rise to question-like interpretations (Kratzer and Shimoyama, 2002; Mascarenhas, 2011), and we have accordingly shown elsewhere that fallacies similar to the ones surveyed in this article can be reproduced with indefinites (Mascarenhas and Koralus, 2015).

5.1.2 Updating via the erotetic principle

The next ingredient is an update rule that implements Part I of the erotetic principle, treating certain premises as questions and others as maximally strong answers to questions in context whenever possible. Suppose now that we have taken on board $\{a\&q, k\&j\}$ as our first premise. We now consider how to implement the idea that the next premise is treated as a maximally strong answer to $\{a\&q, k\&j\}$ (i.e., as a maximally strong answer to the question, "am I in an ace and queen situation or in a king and jack situation?"). We propose that we keep all of those alternatives in the question that have the most in common with the answer. In other words, we keep all alternatives such that no other alternatives have more in common with the answer.[3] The premise we are treating as a answer is $\{a\}$, so this means eliminating the alternative $\{k\&j\}$ from $\{a\&q, k\&j\}$. We are then left with $\{a\&q\}$, at which point a simple rule amounting to conjunct-simplification applies, targeting q, and confirming that the fallacious conclusion follows from the "answered" question.

This simple procedure readily explains the results from experiments 3 and 4 as well (we analyze the order effect of experiment 2 shortly). In experiment 3, the first premise of the illusory inference was (of the shape) $\{a\&q, k\&j\}$, and the second premise $\{k, k\&x\}$. Here, the alternative in the first premise with the fewest atomic proposition in common with the answer is $\{a\&q\}$. Accordingly, the theory eliminates this alternative from the workspace, leaving us with $\{k\&j\}$, from which the observed fallacious inference j follows as before.[4] The premises in example 4 are more complex, but the explanation follows similar lines. The three premises of our sample target in (5) (on page 10) are interpreted as follows.

(13) $\{a\&j\&q, 8\&x\&2, a\}$

[3] A technical remark is in order, aimed at readers familiar with Koralus and Mascarenhas (2013). In light of the data of experiment 4, two revisions to the formalism of Koralus and Mascarenhas (2013) need to be made, replacing the definition of Question-update (Definition 9 on page 332 in the 2013 article), and a resulting simplification of the general update procedure (Definition 11 on page 334 in the 2013 article). The rest of the formalism stays the same and the derivations of reasoning problems discussed in that article are unaffected by these two changes.

Definition 1 (Q-update) *Let a question* Γ *and an answer* Δ *be given. The Q-update of* Γ *with* Δ *is defined as follows.*

$$\langle \Gamma, B, i\rangle[\Delta]^{Q} = \langle \{\gamma \in \Gamma : (\neg\exists\gamma' \in \Gamma) |(\bigsqcap \Delta) \sqcap \gamma'| > |(\bigsqcap \Delta) \sqcap \gamma|\}, B, i\rangle$$

What Q-update does is treat Δ as a maximally strong answer to Γ, keeping all of those alternatives in Γ such that no other alternatives in Γ have more in common with all alternatives in Δ.

Definition 2 (Update) *Let* Γ *be a question and* Δ *an answer. The update of* Γ *with* Δ *is* $\langle \Gamma, B, i\rangle[\Delta]^{Up} = \langle \Gamma, B, i\rangle[\Delta]^{Q}[\Delta]^{C}$.

[4] Strictly speaking, the information that there might be a x is also added, yielding $\{k\&j, k\&j\&x\}$. The theory's rule of conjunct simplification applies to every alternative under consideration, so that we are still left with the fallacious conclusion. We omit these details from the main text to simplify exposition, but the procedure is precisely defined in the cited article.

$\{a \& j \& 8 \& x\}$

$\{\neg q\}$

After update with the second premise, we will be left with only the first two alternatives of premise 1, for they each share two atomic propositions with premise 2. The third alternative of premise 1 shares only one alternative with premise 2, so it is eliminated. Then, premise 3, being contradictory with the first surviving alternative, leaves with the second alternative, namely $\{8 \& x \& 2\}$. From here, the observed fallacious conclusion that there is a 2 follows immediately.

5.2 Dynamic updating of mental model discourses

The erotetic theory is dynamic. That is, the operations that update the workspace of reasoning with the mental representations of successive premises are sensitive to the order of those premises. This follows from the natural dynamics of question asking and answering. In particular, the procedure that implements Part I of the erotetic principle will always take a *new* premise being processed to be an answer to a question that was processed *earlier*. The procedure itself cannot reverse this order and interpret and old premise as an answer to a new question. This accounts for the order effect in experiment 2. To get a fallacious inference in the reversed case discussed in experiment 2, the reasoner would have to mentally take an extra step to change the order of the premises.

It should be noted that besides treating information as questions and answers, our update rule also allows for cases in which we simply accumulate information, as when we are given successive categorical statements.

5.2.1 Simple deduction rule

Reasoning is not just a matter of update. Once reasoners hear and process each premise, they must then be able to perform simple transformations on the resulting mental model, to check what follows. We assume that there is a rule of disjunct simplification, validating the inference $(p \wedge q) \vee r \models p \vee r$. This rule for disjunct simplification includes conjunction simplification as a special case, as the reader can see and is explained in more detail in Koralus and Mascarenhas (2013).

5.2.2 Default reasoning strategy

We make a simple postulate describing how reasoning problems are approached by default. Namely: when given a reasoning problem with premises $P_0, \ldots, P_n$ and conclusion C, reasoners update a blank mental model discourse with each premise, in the order the premises were given. They may then apply the simple deductive rule, targeting the conclusion C. If the resulting mental model in discourse is identical to C, then the inference is deemed valid. Otherwise, it is deemed invalid. If no target conclusion is given for evaluation, reasoners simply update their mental model discourse with all the premise statements and see what holds in the resulting model. The full description of the default reasoning strategy in Koralus and Mascarenhas (2013) includes a model of how background knowledge can influence reasoning performance.

5.2.3 Eliminating contradictions

The theory takes it that reasoners do not immediately see anything wrong with contradictions. However, there must be a process allowing them to look at the representations they are entertaining and check whether they are consistent or not. This comes at a cost and is not part of default reasoning, but it must be a possibility if we want to account for the successes of our reasoning faculty. We therefore define an operation that filters the mental model in discourse, going over each alternative and eliminating all those that are contradictory. This operation also eliminates double negations as described in Koralus and Mascarenhas (2013).

5.2.4 Expanding possibilities through inquiry

It may seem as though saying that reasoners represent "there is an ace and a queen or there is a jack" as the set of alternative possibilities $\{a\&q, j\}$ means attributing to those reasoners an irrational discarding of information. After all, the disjunctive statement allows for a case in which we have an ace, a jack, but no queen. This case is not one of the alternatives in $\{a\&q, j\}$, so are we not really attributing an interpretive mistake to reasoners? Not so. In fact, there is no loss of information at the interpretive stage at all, if we take it that the alternatives reasoners represent for premise statements are simply the exact verifiers of those premise statements (Fine 2012). As already noted, the set of exact verifiers of a statement is the set of all and only those things that exactly make the statement true. For example, $a\&q$ exactly verifies ($a\&q$ or j), as does j. However, while $a\&j\&\neg q$ verifies ($a\&q$ or j), it does not verify this disjunction *exactly*. An exact verifier does not have any superfluous elements that one could take away from it while keeping it a verifier. We can reduce $a\&j\&\neg q$ to j while still having a verifier, so $a\&j\&\neg q$ cannot be an exact verifier. Now, the key to our account of the possibility of correct reasoning is that we can formally recover a full set of "classical" alternatives from a set of exact verifiers. To harness this fact for the erotetic theory of reasoning, we need an operation that expands the mental model under consideration into one that represents every possibility with respect to some propositional atom. We call this operation "inquire." None of the operations in the erotetic theory of reasoning commit us to fallacious inferences (they are, in a relevant sense, sound in principle), so if we "inquire" enough to represent all alternatives allowed by our premise statements, as we could by inquiring on all propositional atoms that have been mentioned, there is no longer any possibility of fallacious inferences. This is the crucial aspect of the theory allowing for classically sound reasoning and it implements Part II of the erotetic principle. In Koralus and Mascarenhas (2013) we prove as a theorem that inquiring on every propositional atom mentioned before updating with new premises guarantees soundness.

5.3 Conclusion

In sum, we believe that the erotetic theory of reasoning provides a distinctly favorable combination of advantages. Like Bayesian approaches, the erotetic theory seeks to account for both successes and failures of reasoning as flowing from a core idea about the computational *aim* of our cognitive capacity for reasoning, though the perspectives offered on that aim by the two theories differ very significantly. According to the erotetic theory, this aim is to answer questions as directly as possible. Like mental model theory, the erotetic theory puts a theory of sparse representations of premises at the center of the explanation of reasoning failures. We hold that, by default, we only represent those alternative possibilities that correspond to exact verifiers of premise statements. Finally, like mental logic, the erotetic theory is formally specified, which makes it possible to calculate predictions and allowed us to prove as a theorem that sufficient inquiry guarantees sound reasoning. Moreover, the erotetic theory shows that technical work in formal semantics can be brought to bear on psychology. Though we do not have the space to present the more general framework here, the erotetic theory is a step toward a general account of reasoning and decision-making. Besides reasoning, the erotetic theory has also been extended to model decision-making (Koralus, 2016; Koralus and Alfano, 2017), and delusional thinking (Parrott and Koralus, 2015).

Acknowledgements

The research reported here was supported by a research grant from the Laces Trust. We thank Sunny Khemlani and Phil Johnson-Laird for helpful discussions.

References

Atlas, J. (2005). *Logic, Meaning, and Conversation: Semantical Underdeterminacy, Implicature, and Their Interface.*. Oxford: Oxford University Press.

Barrouillet, P., Grosset, N., and Lecas, J. (2000). Conditional reasoning by mental models: chronometric and developmental evidence. *Cognition*, 75(3):237–266.

Braine, M. D. S., Reiser, B. J., and Rumain, B. (1984). Some empirical justification for a theory of natural propositional logic. In Bower, G. H., editor, *The Psychology of Learning and Motivation*, chapter 18, pages 317–371. New York: Academic Press.

Braine, M. D. S. and Rumain, B. (1983). Logical reasoning. In Flavell, J. H. and Markman, E. M., editors, *Handbook of Child Psychology: vol 3, Cognitive Development*, pages 263–339. New York: Wiley.

Evans, J. S. B., Newstead, S. E., and Byrne, R. M. (1993). *Human reasoning: The psychology of deduction*. Psychology Press.

Fine, K. (2012). A difficulty for the possible world analysis of counterfactuals. *Synthese*.

Girotto, V., Mazzocco, A., and Tasso, A. (1997). The effect of premise order in conditional reasoning: a test of the mental model theory. *Cognition*, 63:1–28.

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

Harman, G. (1986). *Change in view: Principles of reasoning*. MIT press Cambridge, MA.

Heim, I. (1982). *The semantics of definite and indefinite noun phrases*. PhD thesis, University of Massachusetts Amherst.

Johnson-Laird, P., Legrenzi, P., and Girotto, V. (2004). How we detect logical inconsistencies. *Current Directions in Psychological Science*, 13(2):41–45.

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

Johnson-Laird, P. N. (2008). Mental models and deductive reasoning. In Rips, L. and Adler, J., editors, *Reasoning: studies in human inference and its foundations*, pages 206–222. Cambridge: Cambridge University Press.

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

Khemlani, S. and Johnson-Laird, P. N. (2009). Disjunctive illusory inferences and how to eliminate them. *Memory and Cognition*, 37(5):615–623.

Khemlani, S. and Johnson-Laird, P. N. (2012). Theories of the syllogism: a meta-analysis. *Psychological Bulletin*, 138:427–457.

Khemlani, S., Orenes, I., and Johnson-Laird, P. N. (2012). Negation: A theory of its meaning, representation, and use. *Journal of Cognitive Psychology*, 24(5):541–559.

Koralus, P. (2012). The open instruction theory of attitude reports and the pragmatics of answers. *Philosopher's Imprint*, 12(14).

Koralus, P. (2016). Decisions, illusory reasons, and erotetic equilibrium. Under review.

Koralus, P., and Alfano, M. (2017). Reasons-based moral judgment and the erotetic theory. In: Bonnefon. J.-F. Tremoliere, B. (Eds.). *Moral Inferences*. Psychology Press

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

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

Mascarenhas, S. (2009). Inquisitive semantics and logic. Master's thesis, ILLC.

Mascarenhas, S. (2014). Formal Semantics and the Psychology of Reasoning Building new bridges and investigating interactions. Doctoral Dissertation, New York University.

Mascarenhas, S. (2011). Licensing by modification: the case of positive polarity pronouns. In Guevara, A. A., Chernilovskaya, A., and Nouwen, R., editors, *Proceedings of* Sinn und Bedeutung *16*, pages 417–429.

Mascarenhas, S. and Koralus, P. (2015). Illusory inferences: disjunctions, indefinites, and the erotetic theory of reasoning. In: Noelle, D. C., Dale, R., Warlaumont, A. S., Yoshimi, J., Matlock, T., Jennings, C. D., and Maglio, P. P. (Eds.) *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

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

Oaksford, M. and Stenning, K. (1992). Reasoning with conditionals containing negated constituents. *Journal of experimental psychology. Learning, memory, and cognition*, 18(4):835–854.

Parrott, M. and Koralus, P. (2015). The erotetic theory of delusional thinking. *Cognitive Neuropsychiatry (forthcoming)*.

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

Sauerland, U. (2004). Scalar implicatures in complex sentences. *Linguistics and Philosophy*, (27):367391.

Spector, B. (2007). Aspects of the pragmatics of plural morphology: On higher-order implicatures. *Presuppositions and implicatures in compositional semantics*, 243-281.

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

Conn.	Result	Reference
Not	Few list all alternatives compatible with negated conjunction	Khemlani et al. (2012)
Not	Most can list what corresponds to negated disjunction	Khemlani et al. (2012)
Not	Easy to list a case that falsifies conditional	Oaksford and Stenning (1992)
—	“Explosion” highly counterintuitive	Harman (1986)
Or	Disjunctive syllogism is harder than disjunctive modus ponens	Rips (1994)
Or	Illusory inferences from disjunction	Walsh and Johnson-Laird (2004)
Or	Control problems with disjunction	Walsh and Johnson-Laird (2004)
Or	Supposition makes some problems easier	Johnson-Laird (2008)
Or	Disjunction introduction is counterintuitive	Braine et al. (1984)
Or/If	Fallacies with conditionals and disjunction	Johnson-Laird (2008)
Or/If	Illusory inferences with conditional embedded in disjunction	Johnson-Laird and Savary (1999)
Or/If	Control problems with conditionals embedded in disjunction	Johnson-Laird and Savary (1999)
If	Modus ponens is extremely easy	Braine and Rumain (1983)
If	Modus ponens easier than modus tollens	Evans et al. (1993)
If	Affirming consequent more rapid than denying antecedent	Barrouillet et al. (2000)
If	Order effects on modus tollens	Girotto et al. (1997)
If	Illusions of consistency with sets of biconditional statements	Johnson-Laird et al. (2004)
If	Control problems for biconditional consistency judgments	Johnson-Laird et al. (2004)
If	Illusory inference from disjunction to conditional	Koralus and Mascarenhas (2013)

Table 5: Some core data on naive reasoning captured by the erotetic theory