An interpretation-based account of illusory inferences from disjunction

Salvador Mascarenhas
New York University
smasc@nyu.edu

Sinn und Bedeutung 18 — Vitoria-Gasteiz

September 13, 2013

1. Introduction

- The capacity to reason and draw inferences is as crucial to everyday thought and conversation as it is to modern science and philosophy.
- But it is notoriously prone to failures (Tversky and Kahneman 1974, among many others).
- These phenomena have for the most part been studied by psychologists, who propose accounts rooting such failures in the **general purpose reasoning mechanisms** themselves.
- In this talk I pursue an alternative route: to explain some failures of reasoning as stemming from **interpretive processes**.

$P_1, \ldots, P_n \vdash C$
a. $\vdash$ is non-classical
b. one or more of $P_1, \ldots, P_n$ have non-obvious interpretations

2. Failures of reasoning

2.1. Compelling fallacies and repugnant validities

- I distinguish two broad ways in which human reasoning can fail.
- **Compelling fallacies** are (classically) **invalid** inference patterns that we often **accept**.

Affirming the consequent — accepted by 77% of subjects (Barrouillet et al., 2000)

P_1: If the card is long then the number is even.
P_2: The number is even.
Conclusion: The card is long.

Illusory inference from disjunction — 80% (Walsh and Johnson-Laird, 2004)

P_1: Either Jane is kneeling by the fire and she is looking at the window or otherwise Mark is standing at the window and he is peering into the garden.
P_2: Jane is kneeling by the fire.
Conclusion: Jane is looking at the window.

(1) Illusory inference from disjunction, schematically:
P_1: $(a \wedge b) \vee (c \wedge d)$
P_2: a
Conclusion: b

- The inference in (1) is invalid: consider a world making $a \wedge \neg b \wedge c \wedge d$ true. This models the premises but falsifies the conclusion.
- **Repugnant validities** are (classically) **valid** inference patterns that we often **reject**.

Disjunction introduction (Braine et al., 1984)

P_1: The card is long.
Conclusion: The card is long or the number is even.

- Today's talk will be about certain compelling fallacies, but a complete account of reasoning failures must explain repugnant validities as well.

2.2. Psychological accounts of reasoning failures

- Theories of reasoning failures from psychology identify the **general purpose reasoning mechanism** as the culprit.

- Four important classes of theories from psychology:
  - Heuristics and biases (Tversky and Kahneman, 1974)
  - Bayesian reasoning (Oaksford and Chater, 1991)
  - Mental logic (Rips, 1994): Our capacity for reasoning is underwritten by tacit natural deduction rules, but proofs are hard and we may be mistaken about what the right rules are.
  - Mental models (Johnson-Laird, 1983; Koralus and Mascarenhas, 2013): Reasoning proceeds by manipulating representations of premises. A combination of the rules used and the nature of the representations is responsible for our successes and failures.
- Mental model theory is the only account of reasoning failures that predicts this fallacy.

Mental model theory account of the illusory inference from disjunction (combining elements from Johnson-Laird (1983) and Koralus and Mascarenhas (2013))

- Reasoners build mental representations (mental models) that verify each of the premises.
- The linguistic form of the premise influences which models are considered: only models that make overtly stated material true and are tacit about everything else.
- Disjunctive premises are represented as sets of alternative mental models.
- P_1 gives rise to a set of two alternative models: a minimal model of $a \wedge b$ and a minimal model of $c \wedge d$.
- **Upon hearing P_2, a, reasoners notice that it is related to the first alternative model for P_1, but not the second.**
- This makes them ignore the second model.
- The combined representation of the premises is therefore only one mental model: $a \wedge b$.
- From here, b follows immediately.

3. AN INTERPRETATION-BASED ACCOUNT OF REASONING FAILURES

- There is a natural alternative to the **reasoning-based** accounts found in psychology.
- On an **interpretation-based account,**
  1. there is nothing in principle non-classical about the human capacity for reasoning,
  2. but the **interpretive processes** are more complex that meets the eye. In other words: the premises do not mean what one might think they mean.
- Accounts in this spirit have been given to some classical fallacies within formal pragmatics. Most notably: Horn (2000) discusses affirming the consequent and denying the antecedent.

(2) P_1: If the card is long then the number is even.
P_1': *Only* if the card is long is the number even.
P_2: The number is even.
Conclusion: The card is long.

- But a general research program to extend these kinds of accounts to the more sophisticated reasoning data found in the psychological literature has not been pursued systematically so far.

3.1. Preview: accounting for the illusory inference from disjunction

- The illusory inference from disjunction follows **classically** if we assume that a classically-tuned reasoning module acts on the pragmatically strengthened meaning of the premises.

(3) Illusory inference from disjunction, schematically:
P_1: $(a \wedge b) \vee (c \wedge d)$
P_2: a
Conclusion: b

(4) Strengthened meaning of (3):
$P_1{}^+$: $(a \wedge b \wedge \neg c \wedge \neg d) \vee (c \wedge d \wedge \neg a \wedge \neg b)$
$P_2{}^+$: a
Conclusion: b

- NB: $P_1{}^+$ in (4) is not an obvious implicature of P_1. Simply assuming that the disjunction in P_1 of (3) is exclusive is **not enough**:

(5) $(a \wedge b \wedge \neg(c \wedge d)) \vee (c \wedge d \wedge \neg(a \wedge b))$
$\leftrightarrow (a \wedge b \wedge (\neg c \vee \neg d)) \vee (c \wedge d \wedge (\neg a \vee \neg b))$

3.2. Central components of an interpretation-based account

1. Commitments with respect to *literal* linguistic content — presumably a unidimensional classical semantics, though nothing in the theory will hinge on this hypothesis, which may well have to be qualified in the end.
2. A mechanism for enriching (strengthening) the literal content in a way that
   (a) assigns to each premise the interpretation required to get the observed reasoning patterns as a product of classical deduction rules,
   (b) can be independently motivated as a plausible interpretation of the premises by purely linguistic criteria, and
   (c) introduces no mischief into extant accounts of enriched, non-literal meaning.
3. Basic commitments about reasoning processes—how do reasoners go about checking whether something follows from a set of premises?

3.3. An implicature-based account of some reasoning failures

3.3.1. Scalar implicatures: a neo-Gricean perspective

- Scalar implicatures are a certain kind of quantity implicatures where the hearer compares the speaker's utterance S to a certain class of statements that the speaker could have made but chose not to: those statements that result from substituting elements of S with members of their *scales*.
  1. Compute the alternatives to S, by replacing scalar lexical items in S with elements of their scales.
  2. Collect those sentences S' that are (1) alternatives to S and (2) stronger than S (that is, $S' \vDash S$ but $S \nvDash S'$). Call this set A.
  3. Compute *primary implicatures*: for each sentence $S' \in A$, "the speaker does not believe (i.e. is not in a position to assert) S'."
  4. Compute *secondary implicatures*: assume that the speaker is opinionated, that is, for every sentence S the speaker either believes S or its negation. It follows by disjunctive syllogism that the primary implicatures can be strengthened from the form "the speaker does not believe S'" to the form "the speaker believes that S' is false."

(6) John or Mary will come to the party.
   a. Primary implicature: the speaker does not believe that both John and Mary will come to the party.
   b. Secondary implicature: the speaker believes that it's not the case that both John and Mary will come to the party.

- But something is missing from (6a): we also want as primary implicatures the propositions that "the speaker does not believe that John will come to the party" and "the speaker does not believe that Mary will come to the party."

3.3.2. Enriching the set of formal alternatives

- Following Katzir (2007), I incorporate an appeal to judgments of complexity of the allowed substitutions, abandoning the lexically stipulated Horn scales.

(7) For two syntactic structures S and S', we say that S' is *no more complex* than S, just in case S' can be derived from S by successive replacements of sub-constituents of S with elements of the *substitution source* for S.

(8) For S a syntactic structure, the substitution source for S in C is the union of the following sets:
   a. the lexicon, and
   b. the sub-constituents of S.

3.3.3. From primary implicatures to secondary implicatures

- Consider the simple disjunctive example discussed above, with the improved theory of formal alternatives.

(9) John or Mary will come to the party
   a. Alternatives: {John and Mary will come to the party, John will come to the party, Mary will come to the party}
   b. Primary implicatures: it is not the case that the speaker believes any of the alternatives in a.

- Katzir's (2007) formal alternatives get us the desired primary implicatures.
- But now we must be precise about which primary implicatures get strengthened to secondary implicatures.
- Following Sauerland (2004), I take it that the strengthening procedure from primary implicatures to secondary implicatures must obey the constraint in (10).

(10) No secondary implicature of a statement S can contradict the literal meaning of S or the primary implicatures of S.

3.3.4. Synthesis

- The theory of scalar implicature that I adopt here can be described as the following procedure.
  1. Compute the alternatives to S that are at most as complex as S (definition in (7)).
  2. Collect those alternatives S' that are (1) alternatives to S and (2) strictly stronger than S. Call this set A.
  3. Compute primary implicatures: for each sentence $S' \in A$, "the speaker does not believe that S'."
  4. Compute secondary implicatures: for each $S' \in A$ such that the negation of S' does not contradict the literal meaning of S or any of the primary implicatures of S, conclude (that the speaker believes) that S' is false.
  5. Call the conjunction of the literal meaning of S together with all of its secondary implicatures the strengthened (exhaustive) meaning of S.
- Finally, we need two simple principles about how pragmatic strengthening figures into reasoning:

(11) Even when interpreting sentences in the absence of a speaker, as in a piece of paper in the context of an experiment, reasoners accomodate the existence of some abstract speaker, the author of the sentences under evaluation.

(12) **Reasoning in the implicature-based account**
Given a sequence of premises $P_0, \ldots, P_n$ and a conclusion C, begin by calculating the strengthened meaning of each premise, getting the sequence $P_0^+, \ldots, P_n^+$. Then, check to see if the conclusion C follows *classically* from $P_0^+, \ldots P_n^+$.

4. THE ILLUSORY INFERENCE FROM DISJUNCTION

(13) P_1: Either Jane is kneeling by the fire and she is looking at the window or otherwise Mark is standing at the window and he is peering into the garden.
P_2: Jane is kneeling by the fire.
Conclusion: Jane is looking at the window.

(14) P_1: $(a \wedge b) \vee (c \wedge d)$
P_2: a
Conclusion: b

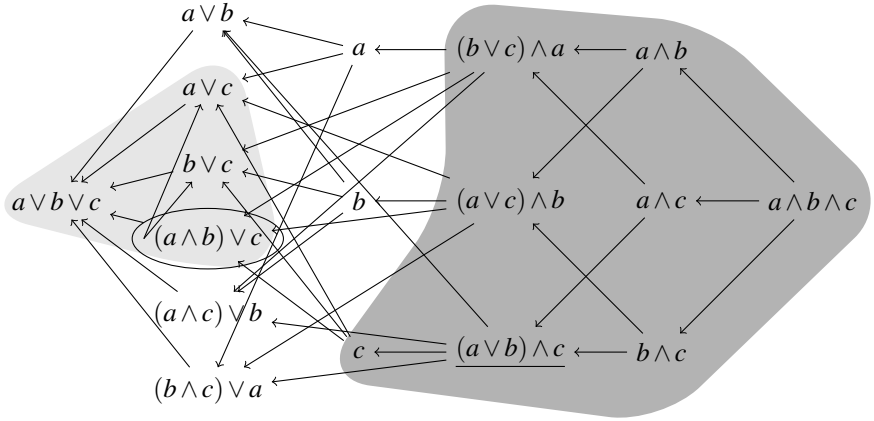

Figure 1: All formal alternatives, up to classical equivalences, for the source $(a \wedge b) \vee c$ (circled in the figure). Arrows between alternatives indicate entailment (transitivity is assumed). The lightly shaded alternatives on the left are weaker than or equivalent to $(a \wedge b) \vee c$. The darker alternatives on the right are strictly stronger than $(a \wedge b) \vee c$.

- To make a fully explicit account possible, I'll consider a variant of (14) with one fewer propositional atom:

(15) P_1: $(a \wedge b) \vee c$
P_2: a
Conclusion: b

- The result I prove in this section was also proved by Spector (2007) in a minimal-models framework, in a context unrelated to the reasoning literature discussed here.
- The first step is to calculate the formal alternatives to P_1 of (15). This is given in Figure 1 (top of this page).
- Each expression in Figure 1 is the result of a licensed substitution according to the adopted theory of formal alternatives. This set is complete, up to certain equivalences we need not worry about.
- Next, we calculate primary implicatures for those alternatives that are strictly stronger than P_1. There are eight:

$\mathbf{c}$	$(a \vee b) \wedge c$
$(\mathbf{b} \vee \mathbf{c}) \wedge \mathbf{a}$	$a \wedge c$
$(\mathbf{a} \vee \mathbf{c}) \wedge \mathbf{b}$	$b \wedge c$
$\mathbf{a} \wedge \mathbf{b}$	$a \wedge b \wedge c$

- The predicted primary implicatures are propositions of the form "the speaker is not in a position to assert φ," for each φ in the set of stronger alternatives above.
- Secondary implicatures: For each of the eight alternatives stronger than P_1, we ask whether we can negate that alternative while not contradicting the literal meaning P_1 or any of the primary implicatures.
- This gives us the set of secondary implicatures in (16).

(16) $\{\neg((a \vee b) \wedge c), \neg(a \wedge c), \neg(b \wedge c), \neg(a \wedge b \wedge c)\}$

- NB: the first secondary implicature in (16) entails all secondary implicatures. We can therefore ignore the remaining three elements of (16).
- Finally, we calculate the strengthened meaning $P_1{}^+$ of P_1, by conjoining the literal meaning P_1 with the secondary implicature $(\neg a \wedge \neg b) \vee \neg c$:

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

- By distributivity of the second conjunct into the first, this is equivalent to

(17) $((a \wedge b) \wedge \underline{((\neg a \wedge \neg b) \vee \neg c)}) \quad \vee \quad (c \wedge \underline{((\neg a \wedge \neg b) \vee \neg c)})$,

- which is in turn equivalent to (18).

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

- Finally, we observe that from the strengthened meaning of the premises, the illusory inference from disjunction is in fact **classically valid**.

(19) $P_1{}^+$: $(a \wedge b \wedge \neg c) \vee (c \wedge \neg a \wedge \neg b)$
$P_2{}^+$: a
Conclusion: b

- This result carries over to the original illusory inference from disjunction, with four propositional atoms.

5. Discussion

- We have an account of the illusory inference from disjunction that explains it as following **classically** from the output of more sophisticated interpretive processes than meet the eye.
- This contrasts with psychological accounts of the same inference that posit non-classical reasoning procedures acting upon the outputs of simplistic interpretive processes.
- There are (at least) two ways in which we can find confirmation for the interpretation-based account.
  1. Look for contexts known to block the required implicature. Acceptance rate for the fallacy should drop considerably.
  2. Look for semantically very similar inference patterns that however have different alternative-sets, so that the required implicature is not predicted to arise. Acceptance rates for these new putative fallacies should be significantly lower than for the original illusory inference.

Quantified illusory inferences

- The propositional illusory inference can be recast with quantifiers doing the job of conjunction or disjunction:

(20) a. Illusory inference from disjunction with universal quantifiers
P_1: Every boy or every girl is coming to the party.
$(P(john) \wedge P(bill)) \vee (P(mary) \wedge P(sue))$
P_2: John is coming to the party.
Q: Does it follow that Bill is coming to the party?

b. with indefinites
P_1: Some student smokes.
$(Stud(j) \wedge Smok(j)) \vee (Stud(m) \wedge Smok(m))$
P_2: John is a student.
Q: Does it follow that John smokes?

Slight modification of the theory of scalar implicature assumed
We now consider alternatives that are *not weaker* than the literal meaning, rather than only those that are *stronger*.

Prediction for universal quantifiers (20a)

- Strengthening of P_1 of (20a):
  (Every boy and no girl) or (every girl and no boy) is coming to the party.
- due to the following alternative:
  Some boy and some girl are coming to the party.

Prediction for indefinites (20b)

- Strengthening for P_1 of (20b):
  Only one student smokes.
- **This is not enough to derive the inference classically.**

- The prediction about indefinites **differs** from the prediction made by Koralus and Mascarenhas (2013). Koralus and Mascarenhas (2013) (and Johnson-Laird and collaborators' mental model theory, if it were general enough) expect (20b) to be a fallacy just like (20a) or the propositional illusory inference from disjunction.
- In an experiment jointly conducted with Philipp Koralus we found that, while (20b) **may be** significantly more accepted than other invalid controls, it is **certainly** significantly **less** accepted than the propositional illusory inference from disjunction — 30% (indefinites) 80% (propositional).
- For Koralus and Mascarenhas (2013) there is no way to account for this difference.

(21) Stimuli samples:

a. P1: Some/a certain student smokes.
P2: John is a student.
C: John smokes.

b. Some/a certain student smokes.
John smokes.
John is a student.

Simplest explanation
the propositional illusory inference is (at least in large part) due to an implicature of the first premise.

- A pilot experiment embedding the P_1 in an *if*-clause showed no drop in acceptance rates...

(22) Illusory inference from disjunction, standard reasoning problem format
P_1: Ann and Bill or Chad and Dan are coming to the party.
P_2: Ann is coming to the party.
Q: Does it follow that Bill is coming to the party?

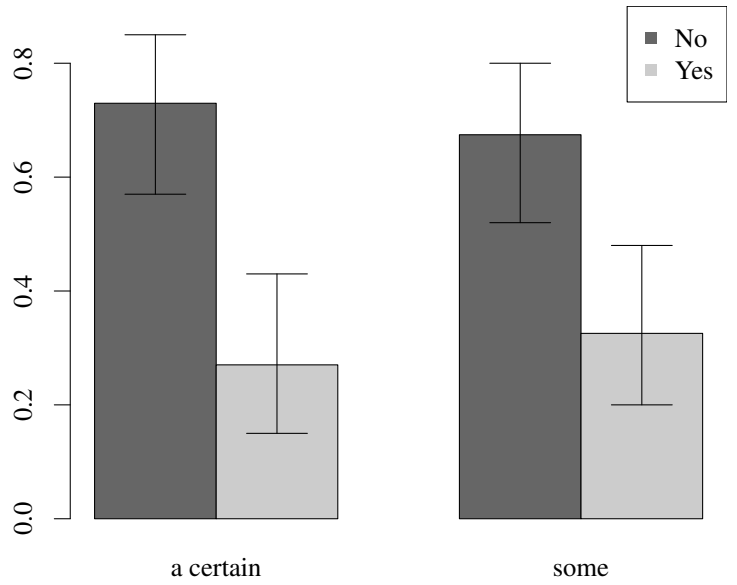

Figure 2: Responses to target inference (percentages) for two kinds of indefinite constructions

(23) Illusory inference from disjunction, conditional format
If Ann and Bill or Chad and Dan are coming to the party, then, if Ann is coming to the party, Bill is also coming to the party.
Q: True or false?

- There are plausible (*post hoc*) reasons for this failure. Sentences like (23) are rather hard to parse, and subjects may have substituted simpler coordination structures for the crucial embedding in (23).

(24) Conditional format from (23)— possible interpretation by subjects
Suppose Ann and Bill or Chad and Dan are coming to the party, and Ann is coming to the party. Is Bill also coming to the party?

6. Conclusion

- I have defined a program for the study of failures of reasoning that roots compelling fallacies in interpretive processes, rather than in the general-purpose reasoning mechanisms themselves.
- I showed that this program can be applied to a class of sophisticated reasoning data from the psychological literature, thus far ignored by the field of formal pragmatics, yielding a natural account that uses only independently motivated interpretive mechanisms.
- I gave empirical evidence that some of the predictions of the interpretation-based theory are borne out.
- Work in progress: finding better ways to test the predictions by blocking the required implicature from the propositional illusory inference.
- This program and this result are of significance to psychology.
- Most scholars of human reasoning do not have a background in linguistics and most linguists do not work on reasoning, so extant theories of reasoning tend not to take advantage of the sophisticated theories of meaning that semanticists have developed over the past forty years.
- Consequently, the difference between general-purpose reasoning and interpretive processes is not well understood.
- Most psychologists would agree that understanding how human reasoning differs from normative logic is an important step toward understanding human reasoning.
- We can only trust our accounts of this intermediate step if we can also trust our understanding of the line between reasoning and interpretation. Without that, the scientists themselves might be falling prey to illusions of human irrationality.
- But linguists should also care about this program.
- In semantics we are interested in the interpretation of linguistic signs, and we study those interpretations partly by inspecting the inferences (entailments, implicatures, presuppositions) validated by utterances.
- The literature on reasoning should be seen by semanticists as a rich repository of inferences, very many of which should in fact be accounted for by our own theories.
- The connection between reasoning and interpretation is a particularly promising domain for beneficial interactions (and hopefully convergence) between psychology and linguistics.

References

Barrouillet, Pierre, Nelly Grosset and Jean-François Lecas (2000). Conditional reasoning by mental models: chronometric and developmental evidence. *Cognition*, 75(3):237–266.

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

Horn, Laurence (2000). From *if* to *iff*: conditional perfection as pragmatic strengthening. *Journal of Pragmatics*, 32:289–326.

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

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

Koralus, Philipp and Salvador Mascarenhas (2013). The erotetic theory of reasoning: bridges between formal semantics and the psychology of deductive inference. forthcoming in Philosophical Perspectives.

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

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

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

Tversky, Amos and Daniel Kahneman (1974). Judgment under uncertainty: heuristics and biases. *Science*, 185:1124–1131.

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