Assessing the role of matching bias in reasoning with disjunctions

Mathias Sablé-Meyer [a] & Salvador Mascarenhas [b]

[a] NeuroSpin, Cognitive Neuroimaging Unit, CEA. Collège de France
[b] Ecole Normale Supérieure, Department of Cognitive Studies, Institut Jean-Nicod

Summary

- On mental models theories, reasoners create mental representations of information, which they manipulate in order to derive new conclusions.
- These theories have been uniquely successful at explaining a class of attractive fallacies involving disjunctions.
- In this article we examine a crucial ingredient of mental models accounts of these illusions, a matching procedure.
- In three experiments:
  - We show that what is explained in terms of low-level matching or overlap in content in these theories must in fact take place at a higher level of cognition.
  - We introduce variants of illusory inferences from disjunction whose acceptance by participants is accurately predicted by their confidence in causal connections that rely on world knowledge.

Background

Illusory Inferences From "Disjunction"

John speaks English and Mary speaks French, or else Bill speaks German.
John speaks English.
Fallacious conclusion: Mary speaks French. (adapt. Walsh & Johns.-Laird, 04)

(a ∧ b) ∨ c
a
Fallacious conclusion: b

- Participants accept this (85%) regardless of how the disjunction is expressed ('or', 'either [...] or', 'or otherwise'; Mascarenhas, 2014)
- This has been shown in 'does [fallacy] follow? [yes/no]' and in 'what if anything follows?' paradigms (Koralus & Mascarenhas, 2018)
- Elements opaquely related to disjunction trigger similar behavior:

SOME pilot writes poems.
John is a pilot.
Concl.: John writes poems. (Mascarenhas & Koralus, 2017)

Miranda MIGHT play the piano and be afraid of spiders.
Miranda plays the piano.
Concl.: Miranda is afraid of spiders. (Mascarenhas & Picat, 2019)

Theories

- The first illusory inferences from disjunction were discovered and analyzed within the **Original Mental Model Theory** (OMMT) (Johnson-Laird, 1983)
- The **Erotetic Theory of Reasoning** (ETR) (Koralus & Mascarenhas, 2013) offers a variant of OMMT that incorporates insights from linguistic semantics and formal logic:
  - Reasoners build mental models that verify each of the premises.
  - Disjunctive premises give rise to **sets of mental models** that **ask a question**: which of these alternatives is the case?
  - Models of the disjunctive premise (the **question**) that do not **match** (overlap with) the model of the second premise (the **answer**) drop from attention.
  - In the example above, reasoners are left with "John speaks English and Mary speaks French," whence the fallacious conclusion "Mary speaks French."
- On **probabilistic accounts** (Oaksford & Chater, 2007), the probability of a putative conclusion conditional on the premises determines whether a fallacy is predicted.
- **However**, assuming equiprobability and independence of the propositions in the first premise, these theories predict the observed fallacious conclusion "John Mary speaks French" **to the same extent** as an unobserved fallacy "Bill speaks German."

References

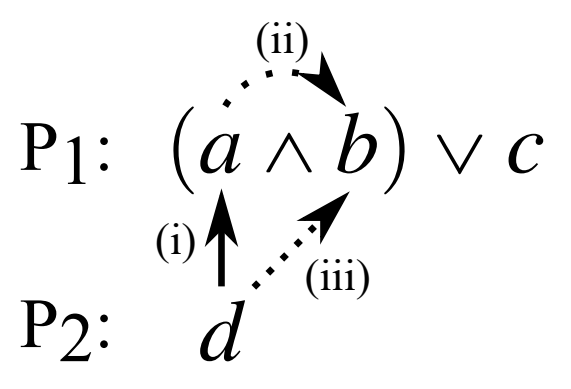
- Cummins, D. D. (1995). Naive theories and causal deduction. *Memory and Cognition*
- Johnson-Laird, P. N. 1983. Mental models: towards a cognitive science of language, inference, and consciousness. *Cambridge: Cambridge University Press.*
- Mascarenhas, S., & Koralus, P. (2017). Illusory inferences with quantifiers, *Thinking and Reasoning*
- Khemlani, S. S. & Byrne, R. M. & Johnson-Laird, P. N. (2018). Facts and Possibilities: A Model-Based Theory of Sentential Reasoning. **Cognitive science**
- Koralus, P., & Mascarenhas, S. (2013). The erotetic theory of reasoning: bridges between formal semantics and the psychology of deductive inference. *Philosophical Perspectives*
- Mascarenhas, S., & Picat, L. (2019). Might as a generator of alternatives, *SALT*
- Oaksford, M., & Chater, N. (2007). Bayesian rationality: The probabilistic approach to human reasoning. *Oxford University Press.*
- Walsh, C. & Johnson-Laird, P. N. (2004). Coreference and reasoning. **Memory and Cognition**

Main Experiment Design

- Goal: **to test the nature of the matching procedure.**
- In order to do this,
  - We used a second premise that **did not match** anything in the first premise exactly (schema on the right),
  - but instead was **causally connected to it**.
  - We measured beliefs about the strength of various causal connections $d \to a$ (schema on the right),
  - We measured the acceptance rate of fallacies that depend on these causal connections.
- By removing superficial elements of the matching step we hoped to observe a **gradient of fallacies tied to world-knowledge**, demonstrating that the low-level matching is not an account of the phenomenon.
- Internal confounds could arise as the fallacy can follow from:

(i) $d \to a$ via the expected indirect matching procedure
(ii) $a \to b$ via a sequence of entailments, $d \to a \to b$
(iii) $d \to b$ directly

- Consequently we had one group of subjects ($N = 153$) rate the strength of the connections (i), (ii), and (iii) in conditional form.
- Another group ($N = 64$) solved a traditional "does [fallacy] follow? [y/n]" task with the same materials.



(a ∧ b) ∨ c
d
Fallacious conclusion: b

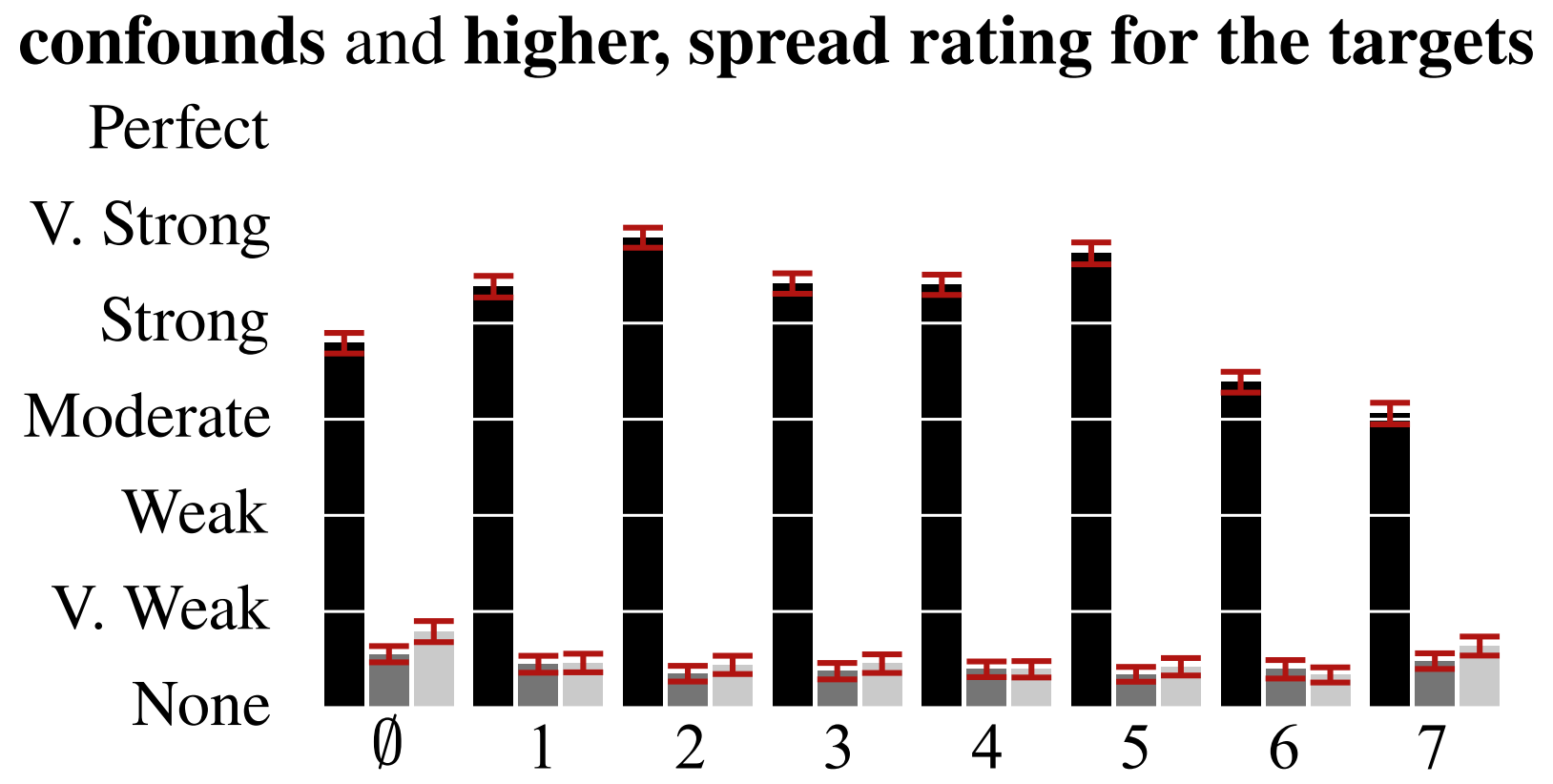
- We adapted materials from work on causal connections and conditional reasoning (Cummins, 1995) so that we would have:
  - High ranking, high variance $d \to a$ connection (**black** on the graph below)
  - As low as possible connections for (ii, **dark grey**) and (iii, **light grey**)

1. If the brake was depressed, then the car slowed down.
2. If Mary jumped into the swimming pool, then Mary got wet.
3. If the trigger was pulled, then the gun fired.
4. If Larry grasped the glass with his bare hands, then Larry left fingerprints on his glass.
5. If the gong was struck, then the gong sounded.
6. If John studied hard, then John did well on the test.
7. If the apples were ripe, then the apples fell from the tree.

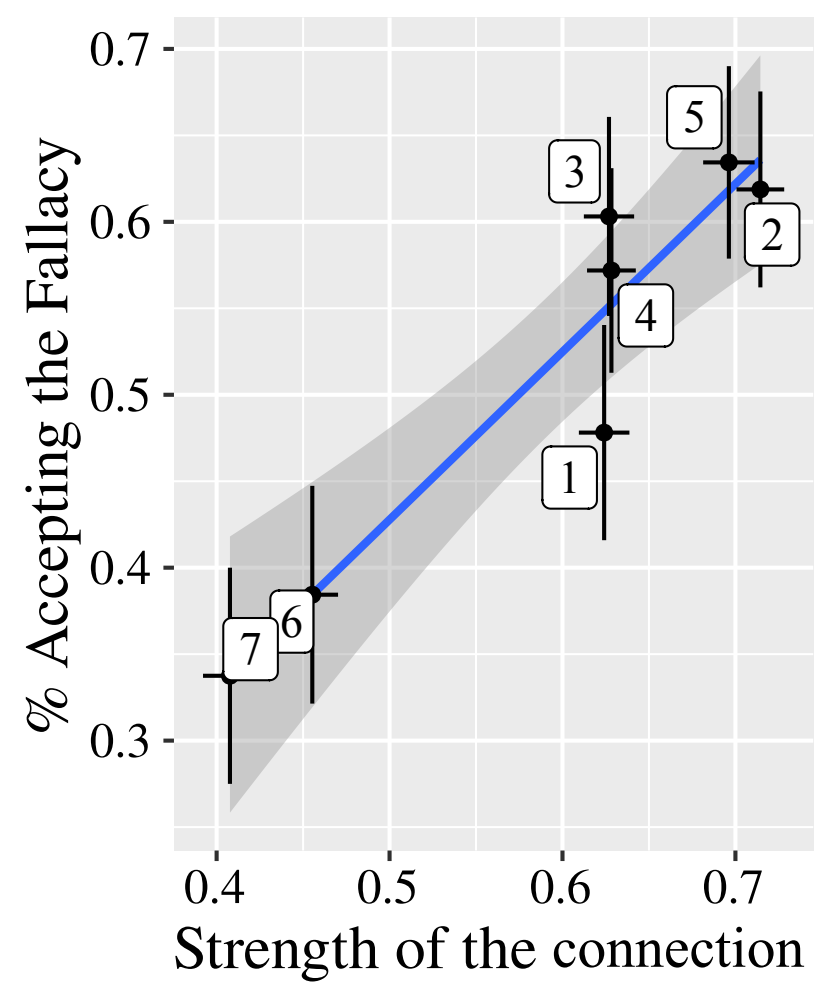
∅. If fertilizer was put on the plants, then the plants grew quickly

Results

- Results are shown below. ∅ was removed from the fallacy experiment as it was driving all the effect of a significant difference for (ii) and (iii). Notice the **low ratings on confounds** and **higher, spread rating for the targets**



- There is a **tight correlation** between the measure of the acceptance rate and the rating of the d to a causal link, as shown on the graph below



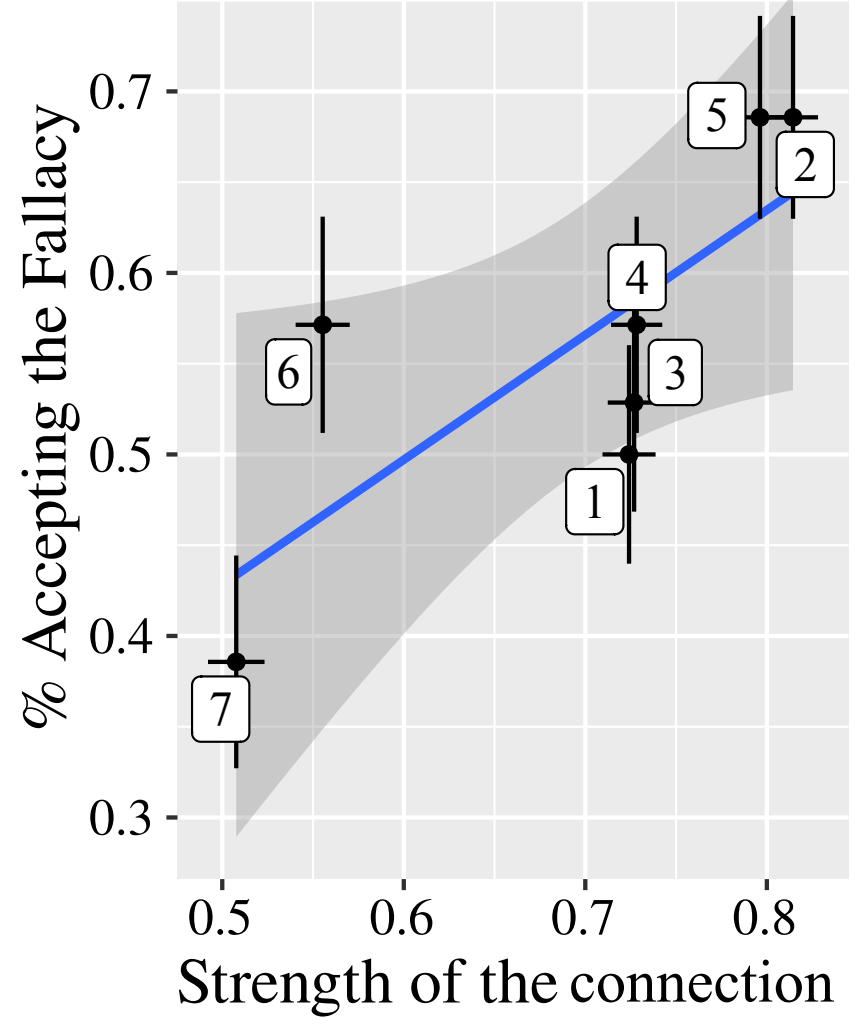
- A linear model of the effect of the rating on the fallacy **successfully predicts the acceptance rate**, and after scaling both variables to [0,1] the slope is not significantly different from 1 with $R^2 = 0.9$ and $p < 0.005$
- Taking the control questions into account, it appears the **the more rational or attentive the participant, the stronger the effect**

Exp 2 — a variation

- In experiment 2 we used the rating collected in the first experiment to check for an indirect fallacious inference of another nature:

$(b \wedge d) \vee c$
b
Does it follow that a? (where independently d causally connects to a)

- The first step allows for low-level **matching**,
- Once that is accomplished, **causal connections are required** to get the conclusion indirectly from the right-hand side of the conjunction.
- The plot on the right shows the correlation between the acceptance rate and the independent measures on the targets.
- A linear model of the effect of the rating on the fallacy **successfully predicts the acceptance rate**. After scaling both variables to [0,1] the slope is 0.7 with $R^2 = 0.58$ and $p = 0.046$
- In addition to the p value being less convincing, participant's score on controls does not significantly predict their acceptance rate — unlike in experiment 1



Exp 1 & 2 — Discussion

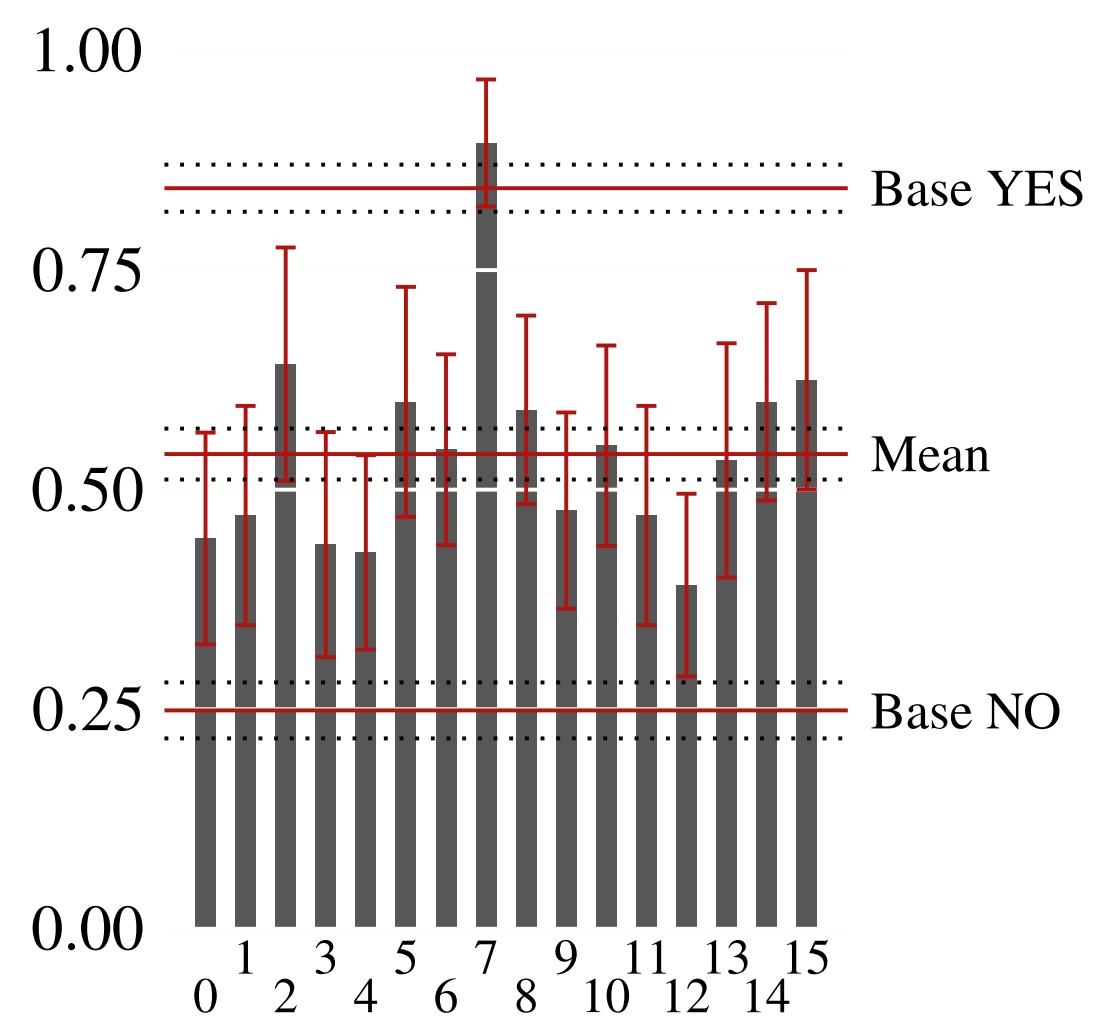
- People accepted the fallacious conclusion **even in the absence of a clear "matching" procedure**
  - The extent to which participants accept the fallacy **strongly correlates with an independent measure** of the perceived strength of the connection
  - This occurs whether the indirectness is required in lieu of the matching procedure, or as a post-matching step
- Classical illusory inferences are the special case where $d = a$. Since they typically yield about 85% success rate, one might expect a ceiling effect
- This confirms that the procedure is **part of a rich reasoning process with multiple moving parts**, as not a low-level shortcut to an immediate conclusion.

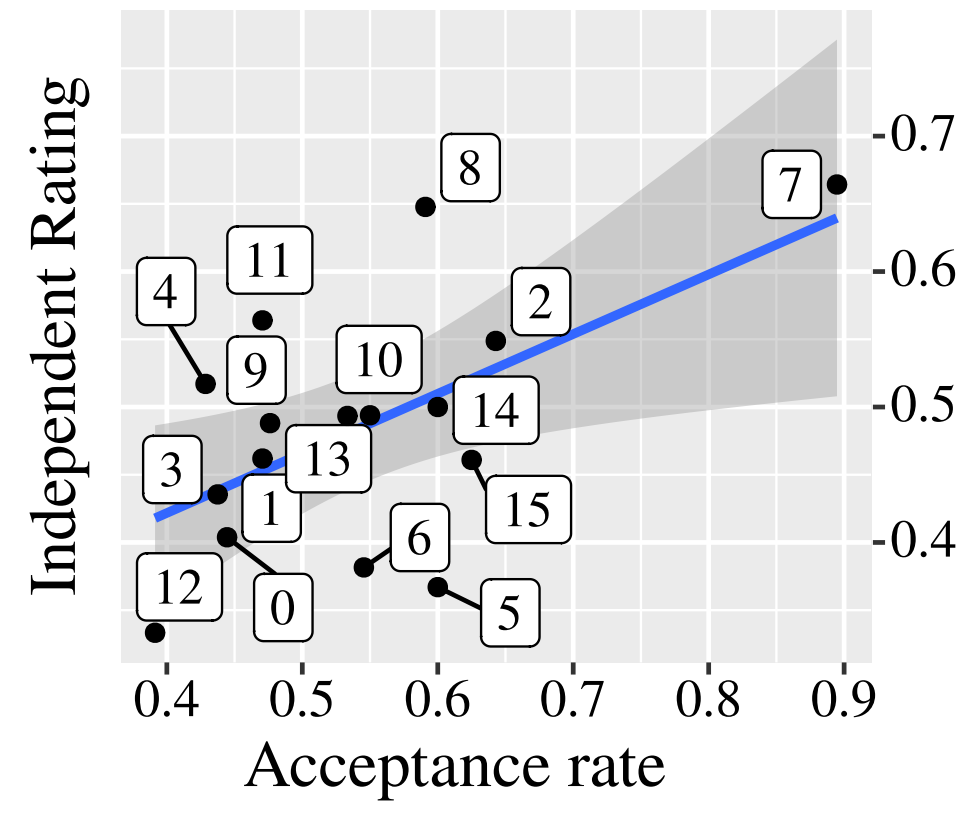
Exp 3 — Pilot

Every guest at this party either owns a car or lives in New York.
Chloe is at the party and lives in the US.
Does it follow that Chloe lives in New York?

- We explored one more way in which premises can be related without matching
  - Using set inclusion and supersets as cues
  - With reversed sets inclusion serving as controls
- Two independent populations
  - A first group ($N = 80$) rated the strength of set inclusions through a task like the following: "This party is for all people who X. Marie is at the party and Y. How confident are you that Marie is allowed at the party?" where $X \subset Y$ or the converse.
  - A second group ($N = 40$) solved "does [fallacy] follow? [y/n]" with the same materials and the structure given in the example above

Exp 3 — results & discussion





- Participants' responses to targets show that the fallacy is more attractive than the invalid ("no") controls. They performed very well on valid ("yes") controls
- This appears to come from a **question-answer pattern** similar to that of experiments 1 and 2, **despite the absence of a clear matching strategy** and necessarily using external beliefs
- The correlation is significant ($p = 0.019$) but does not explain the data as well as in the previous cases ($R^2 = 0.33$)

Conclusion

- We have shown that illusory inferences from disjunction are in fact a sophisticated phenomenon that requires recruiting world knowledge rather than relying on low-level matching strategies
- The Revised Mental Model Theory (RMMT) incorporates world knowledge and system 1 and 2 processes. However, the proposed account of reasoning only gets these disjunction fallacies to the extent that:
  - In system 1, it also gets **the unobserved conclusion** c
  - In system 2, it does a **fully exhaustive pragmatic reading of the first premise**, which cannot account for the whole phenomenon such as the examples with 'might' (Mascarenhas & Picat, 2019)
- We posit that **a Bayesian confirmation-theoretic implementation** of OMMT or the erotetic theory accounts for the examples above. We propose that, when entertaining competing alternatives, reasoners engage in a form of **hypothesis testing**. We are currently developing such a theory
- This view **connects illusory inferences from disjunction to the conjunction fallacy** for which confirmation-theoretic accounts in a similar vein have been proposed.

(b ∧ f) ∨ b
⟨description fitting f⟩
Fallacy: b ∧ f

LaTeX TikZposter