Good Explanations Fit Prior Knowledge

(a poster secretly about pronouns)
(and possible worlds reasoning)

Joshua K Hartshorne[a], Tobias Gerstenberg[b], Noah Goodman[b]
[a]MGH Institute of Health Professions [b]Stanford University

Introduction

Work on pronouns has highlighted role of explanations:

1. Albert frightened Bart **so** he… [he = Bart]
2. Albert frightened Bart **because** he… [he = Albert]

But explanations come in multiple types**:**

3. The grass is wet because it rained last night. [causal]
4. It rained last night because the grass is wet. [existential]

compare with:

5. Albert frightened Bart because he is scary. [causal]
6. Albert frightened Bart because he looks scared. [epistemic]

Hypothesis

Intuition:

- *Because* sentences are ambiguous because causal and existential meanings.
- Wet grass is good evidence for rain, but poor cause
- Rain is good cause of wet grass and good evidence (cf. 3)
- But if speaker wanted to convey epistemic explanation, could have said:
  - "*I believe* it rained last night because the grass is wet*.*"

Hypothesis**:** Listeners choose interpretation in proportion to plausibility.

Experimental Data

Created 16 sentence pairs like (3) and (4):

- Alfred is strong because Alfred beat Bart at tug-of-war
- Alfred beat Bart at tug-of-war because Bart is strong

Created explicit epistemic versions:

- I think Alfred is strong because Alfred beat Bart at tug-of-war
- I think Alfred beat Bart at tug-of-war because Bart is strong

72 adults rated relative probability of each interpretation for each of the 64 sentences

Model

Rational observer:

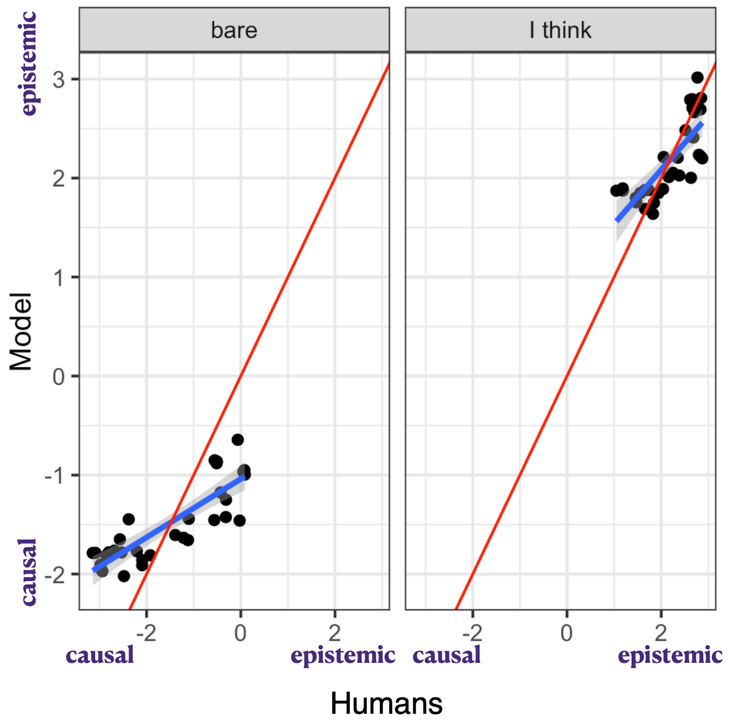
$$P(\text{meaning}|\text{utterance}) \propto P(\text{utterance}|\text{meaning}) * P(\text{meaning})$$

P(meaning) obtained from MTurk.

- "How likely is it that rain would cause the grass to be wet?"
- "How likely is it that observing rain would cause someone to believe the grass is wet?"
- "How likely is it that wet grass would cause rain?"
- "How likely is it that observing wet grass would cause someone to believe it rained?"

Fit two values for P(utterance|meaning): one for "I think" and one for non-"I think" sentences

Results and Discussion



Major results

- Model fit: r = .95, 95% CI [.92, .97], t(62) = 24.16, p < .001
- Epistemic meaning preferred for "I think" sentences
- Causal meaning preferred for non-"I think" sentences
- Controlling for "I think", strong effect of plausibility

Discussion

- Interpretation follows plausibility: P(meaning)
- Plausibility was judged by humans but can be derived from possible worlds model (stay tuned)

Future Directions

- Derive P(meaning) from possible worlds model
- Do different explanation types affect pronoun interpretation?
- Confirm in other languages

Bibliography

- Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. Psychological review, 128(5), 936.
- Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. Psychological science, 17(9), 767-773.
- Hartshorne, J. K., O'Donnell, T. J., & Tenenbaum, J. B. (2015). The causes and consequences explicit in verbs. Language, cognition and neuroscience, 30(6), 716-734.
- Kehler, A., & Kehler, A. (2002). Coherence, reference, and the theory of grammar (Vol. 380). Stanford, CA: CSLI publications.