Winograd-schema without world knowledge: Minimal post-pronoun semantics modulate the implicit causality pronoun bias

Amanda Rose Yuile[1] (ayuile@mghihp.edu) & Joshua K. Hartshorne[1]
1. MGH Institute of Health Professions

Introduction. In causally dependent clauses, pronoun interpretation is systematically affected by the main verb. For example, comprehenders typically resolve pronouns to the cause (if any) entailed by the verb (see 1-2) – a phenomenon termed *implicit causality* [1].
(1) Sally feared Mary because she… (2) Sally frightened Mary because she…
These effects can be overridden by material coming after the pronoun, as shown in the classic example from [2]:
(3) a. The city council denied the demonstrators a permit because they feared violence.
b. The city council denied the demonstrators a permit because they advocated violence.
In (3), pronoun interpretation depends on inferences drawn from detailed world knowledge that arrives *after* the pronoun. In contrast, pronoun interpretation in (1-2) reflects a probabilistic bias based on abstract representations of the linguistic context encountered *before* the pronoun. These probabilistic biases are often attributed to predictive processing, in which comprehenders rapidly extract information as sentences unfold (see 3). While compatible with post-pronoun effects, this view does not clearly predict that later-arriving information should systematically override earlier predictions. This raises a key question: Does post-pronoun information guide interpretation only when it provides rich event-level semantics, or can more abstract post-pronoun semantic cues also influence pronoun comprehension?

Some initial hints come from an incidental finding reported in [4]. In two studies examining the implicit causality pronoun bias [4,5], [4] observed a shift toward object interpretations even though it tested hundreds of the same verbs as [4]. The key methodological difference was that [5] ended sentences like (1-2) with a novel **verb** (e.g., *because she daxed*), whereas [4] ended them with a novel **noun** (e.g., *because she was a dax*). Though neither ending has much *specific* meaning (compare with (4)), they do differ on an abstract level (event vs. kind).

Methods. To test this systematically, we asked 319 English-speaking participants read 32 implicit causality sentence fragments such as: *Kimberly feared Brian because*… Each sentence fragment included one male and one female character, and one of 501 transitive verbs (drawn from 7 Levin verb classes [6,7]). On each trial, participants were asked to continue the sentence using one of two options. There were four continuation conditions: **baseline condition** (*he…* / *she…*), **noun condition** (*he was a dax.* / *she was a dax.*), **intransitive verb condition** (*he daxed.* / *she daxed*), and **transitive verb condition** (*he daxed him.* / *she daxed her.*). The choice of pronoun gender indicated the reference bias. All the usual things were randomized across participants and items.

Results. We used a logistic mixed effects regression with random intercepts for verb and condition (random slopes were computationally intractable). Compared with the baseline condition, the noun condition was significantly more subject-biased ($p = .026$) and the verb conditions were significantly more object-biased (intransitive: $p < .001$; transitive: $p = .007$). Results are depicted in Figure 1.

Conclusion. The form of the predicate for which the pronoun is the subject affected pronoun interpretation, even in the absence of detailed lexical or world knowledge. That is, relative to a baseline condition with no post-pronoun material, predicative nouns led to more subject pronoun interpretations, while verbal predicates led to more object pronoun interpretations – despite the nouns and verbs having no known meaning. We will discuss theoretical implications of these findings for the processing of pronouns (i.e., whether probabilistic pronoun biases vs. final (fixed) interpretation rely on the same or different mechanisms).

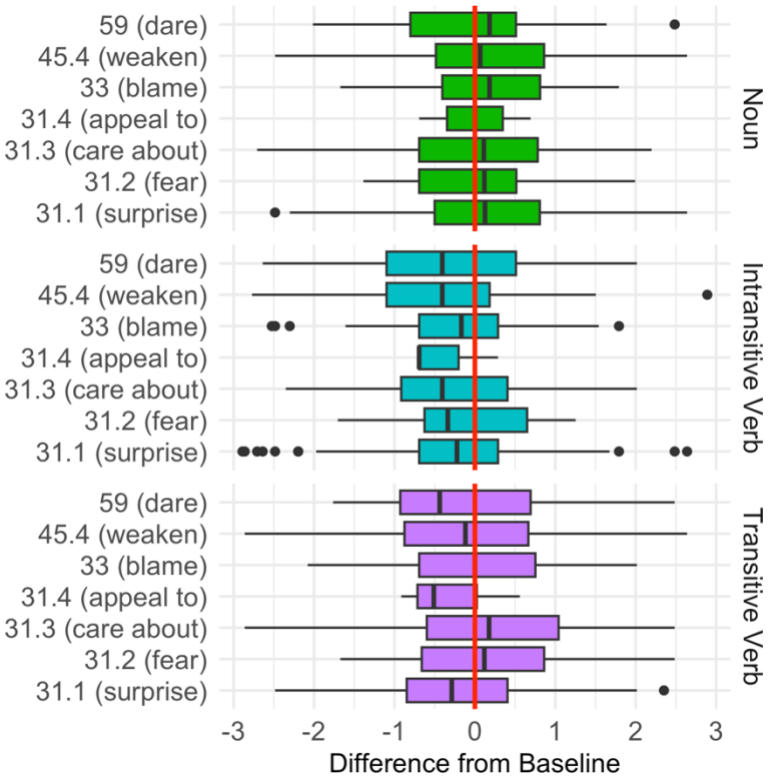

Figure 1. Pronoun interpretation biases for each verb class by condition. Biases are depicted as the difference from baseline in empirical log odds of choosing the subject. The vertical red line indicates the baseline bias.

References

1. Garvey, C., & Caramazza, A. (1974*).* Implicit causality in verbs. *Linguistic Inquiry, 5*(3), 459-464. Retrieved from https://www.jstor.org/stable/4177835
2. Winograd, T. (1972). Understanding natural language. *Cognitive Psychology, 3*(1), 1-191. DOI: 10.1016/0010-0285(72)90002-3
3. Arnold, J.E. (2023). Pronoun comprehension. In The Routledge Handbook of Pronouns (pp. 120-134). Routledge.
4. 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. DOI: 10.1080/23273798.2015.1008524
5. Hartshorne, J.K. & Snedeker, J. (2013). Verb argument structure predicts implicit causality: The advantages of finer-grained semantics. *Language and Cognitive Processes, 28*(10), 1474-1508. DOI: 10.1080/01690965.2012.689305
6. Levin, B. (1993). *English verb classes and alternations: A preliminary investigation. Chicago: University of Chicago Press.*
7. Kipper, K., Korhonen, A., Ryan, N., & Palmer, M. (2006). Extending verb net with novel verb classes. *LREC,* 1027–1032.