A probabilistic, goal-sensitive model of normative rule interpretation

Brandon Waldon (with Cleo Condoravdi, Beth Levin, and Judith Degen)

Sinn und Bedeutung (SuB) 27 Companion Handout
September 15, 2022

This analysis extends the the Rational Speech Act (RSA) framework of modeling pragmatic interpretation (Frank & Goodman, 2012); it takes particular inspiration from RSA analyses whereby listeners jointly infer speaker intended meanings and speaker non-communicative goals. (See e.g. Kao, Wu, Bergen, & Goodman, 2014 for an application to hyperbolic language; Kao, Bergen, & Goodman, 2014 for metaphor; Burnett, 2019 for social meaning and joint inference re: multiple social goals). On our account, interpretation of normative rules is sensitive to contextual evidence as to the rule-maker's intended policy goal (e.g. limiting pollution in the case of *No vehicles allowed*). We call our proposed model the **the goal-sensitive model:**

$$L_1(obj, goal|rule)^1 \propto S(rule|obj, goal)^2 \cdot P_G(goal|rule)^3 \cdot P_{CAT}(obj|rule)^4$$

[1] $L_1(obj, goal|rule)$: Joint posterior probability that object *obj* is prohibited and signaler has particular policy *goal*, given observation of a *rule*.

(Other possible prohibition state: $\neg obj$ [*obj* is not prohibited])

[2] $S(rule|obj, goal)$: Signaler prob. of producing *rule*, given *goal* and intention to prohibit *obj*.

$$\propto e^{\alpha \cdot ln(U(goal,obj) - cost(rule))}$$

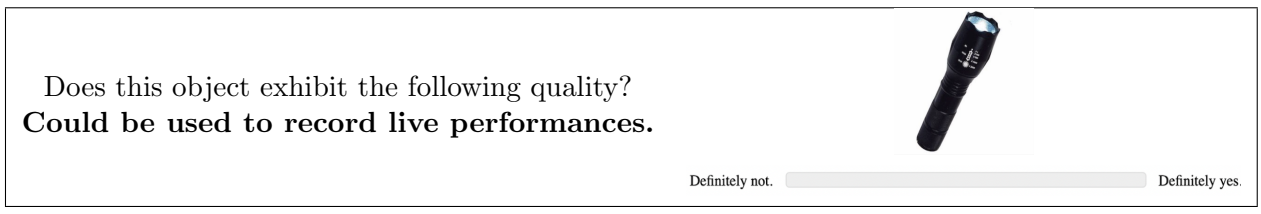
... where α is an optimality parameter, *cost* is a cost function on signal choices, and $U(goal, obj)$:

- Tracks utility of prohibiting *obj* given a policy *goal*.
- Outputs on the interval [0,1] (the greater the extent to which prohibiting *obj* advances *goal*, the higher the output value).

$$U(goal, \neg obj) = 1 - U(goal, obj)$$

(Failing to prohibit objects that would advance *goal* is low utility).

- Is parameterized via a norming study in which objects in the experiment are normed for goal-relevant features, e.g.:



[3] $P_G(goal|rule)$: Prior over policy goals (given observation of *rule*). Parameterized via a norming study in which policy goals are normed for a priori plausibility, e.g.:

The managers of a theater are concerned that audience members might try to record performances and distribute pirate recordings online.

How plausible is it that the **motivation** above could have given rise to the rule below?

No electronic devices are allowed in the theater.

Highly implausible. Highly plausible.

[4] $P_{CAT}(obj|rule)$: Prior beliefs that *obj* is prohibited, given lexical content of *rule*. (In particular, prior beliefs about the membership of *obj* in the category denoted by an artifact noun that features in the rule).

$$P_{CAT}(\neg obj|rule) = 1 - P_{CAT}(obj|rule)$$

Is this object an electronic device?

Definitely not. Definitely yes.

The goal-insensitive baseline model:

$$L_1(obj|rule) = P_{CAT}(obj|rule)$$

For the purposes of model comparison, we assume:

- A space of possible messages that includes a scene's featured *rule* and an alternative, *silence* (following Lassiter & Goodman, 2013) with the following properties:
  - $S(silence|obj, goal) = S(silence|\neg obj, goal)$ for any *obj* and any *goal*.
  - For any $goal_1$, $goal_2$ in the space of possible policy goals, $P_G(goal_1|silence) = P_G(goal_2|silence)$
  - $P_{CAT}(obj|silence) = P_{CAT}(\neg obj|silence)$ for any *obj*
- For trials in which participants are exposed to an explicit policy goal $goal_x$, $P_G(goal_x|rule) = 1$.
- $cost(rule) = cost(silence) = 0$

Model comparison was conducted in WebPPL (Goodman & Stuhlmüller, 2014) from a uniform prior over model architectures (goal-sensitive vs. goal-insensitive) and a uniform prior over α values.

References

Burnett, H. (2019). Signalling games, sociolinguistic variation and the construction of style. *Linguistics and Philosophy*, *42*(5), 419–450.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Goodman, N. D., & Stuhlmüller, A. (2014). *The Design and Implementation of Probabilistic Programming Languages.* `http://dippl.org`.

Kao, J., Bergen, L., & Goodman, N. (2014). Formalizing the pragmatics of metaphor understanding. In *Proceedings of cog sci* (Vol. 36).

Kao, J., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, *111*(33), 12002–12007.

Lassiter, D., & Goodman, N. D. (2013). Context, scale structure, and statistics in the interpretation of positive-form adjectives. In *Semantics and linguistic theory* (Vol. 23, pp. 587–610).