Title

Modeling transfer of high-order uncertain information

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 39(0)

Authors

Herbstritt, Michele
Franke, Michael

Publication Date

2017

Peer reviewed

Modeling transfer of high-order uncertain information

Michele Herbstritt (michele.herbstritt@gmail.com)
Michael Franke (mchfranke@gmail.com)
Department of Linguistics, University of Tübingen
Wilhelmstraße 19, 72070 Tübingen, Germany

Abstract

Complex uncertainty expressions such as *probably likely* and *certainly possible* naturally occur in everyday conversations. However, they received much less attention in the literature than simple ones. We propose a probabilistic model of the use and interpretation of complex uncertainty expressions based on the assumption that their predominant function is to communicate factual information about the world, and that further layers of uncertainty are pragmatically inferred. We collected empirical data on the use and interpretation of these expressions and use it for detailed model criticism.

Keywords: uncertainty; probability; experimental pragmatics; computational modeling

Introduction

One of the main goals of human linguistic interactions is the exchange of information. However, the information that we want to exchange can be uncertain: we often talk about things that we do not know for sure. As a consequence, it should not surprise us that human languages are equipped with so called "uncertainty expressions" such as epistemic modals (*possible*, *might*) and probability expressions (*probably*, *likely*).

Simple uncertainty expressions have been extensively investigated in psychology (Beyth-Marom, 1982; Teigen, 1988; Windschitl & Wells, 1996, 1998) and formal linguistics (Kratzer, 1991; Yalcin, 2010; Egan & Weatherson, 2011; Lassiter, 2011), where some consensus has recently emerged about the advantage of adopting a formal semantics that uses probability measures (contra the purely qualitative semantics à la Kratzer). Herbstritt and Franke (2016) empirically investigated the production of simple uncertainty expressions (*probably*, *possibly*) and propose a pragmatic model of their production. This paper substantially extends the scope of that work: here we investigate complex (or nested) expressions such as *probably likely* and *certainly possible* and we model both their production and interpretation in a conversation.

Complex uncertainty expressions have received much less attention in the literature.[1] Indeed, many foundational issues arise in the attempt to formalize a model of their use and interpretation. Most pressingly are two interrelated concerns: (i) what is the semantic meaning of a complex uncertainty expression? and (ii) what is the communicative goal of a complex uncertainty expression, i.e. what is the pragmatic purpose of communication? In this paper we present a first model that commits itself to what are arguably the most natural answers to (i) and (ii) from the point of view of formal semantics (Swanson, 2006; Moss, 2015) and a rational analysis of communicative practices as efficient transfer of information about the world (Anderson, 1990, 1991).

[1] A recent exception is (Moss, 2015).

This approach enables a straightforward regular and compositional treatment of the meaning of uncertainty expressions: simple and complex uncertainty expressions denote sets of probability distributions over the state space that represents the possible ways in which the world can be. The meanings of simple expressions are always singleton sets. The meanings of complex expressions are derived compositionally in terms of the simple ones and in general they contain more than one distribution (see details below). If we model agents' uncertain beliefs about the world as (sets of) probability distributions over the same state space, then the meaning of a simple or complex uncertainty expression can be seen as a collection of ways to update the agents' beliefs. Figure 1 displays an intuitive representation of this idea.

We incorporate this idea in a probabilistic pragmatic model of language production and interpretation based on the *Rational Speech Acts* (RSA) model (Frank & Goodman, 2012). In particular our model can be seen as a conservative generalization of the RSA model proposed by Goodman and Stuhlmüller (2013). The key innovation of our model is to treat uncertain beliefs of agents (and thus the communicative effect of messages) as sets of probability distributions, hence more fine grained than in the usual approach.

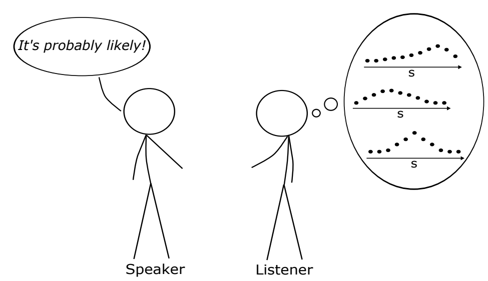

Figure 1: Listener's beliefs as complex uncertainty state. Each probability distribution in the listener's beliefs is compatible with the literal meaning of the received message.

The details of the model are spelled out in the next section. In the following section we report on two experiments designed to collect human data about production and interpretation of complex uncertainty expressions. Finally, the predictions of the model are evaluated against experimental data with Bayesian inference and model criticism.

Pragmatic model

Setup We want to model communication in situations of what we call high-order uncertainty. To illustrate, imagine an urn containing 10 balls of two different colors (e.g., red and blue). The universe of the discourse, the set of possible states of affairs, can be modeled as the set of natural numbers $S = \{0, \dots, 10\}$ where each $s \in S$ is a possible quantity of red balls in the urn. The ratio $s/10$ expresses the objective chance that a randomly drawn ball will be red, and represents first-order uncertainty: even if we know the objective chance, we are uncertain about the color of a randomly drawn ball. The second, high-order level of uncertainty comes into play when we are uncertain about the objective chance too. We model agents that do not have direct access to the objective chance. Instead, one agent (the speaker) can draw a certain number of balls from the urn (referred to as the "access" and denoted with a) and look at them. The set of possible access values is $A = \{1, \dots, 10\}$. The number of red balls among the accessed ones is referred to as the "observation" and denoted o, $0 \leq o \leq a$. We assume that the communication is about the content of the urn: after her observation, the speaker puts all balls back in the urn and makes a prediction about the color of a randomly drawn ball (see Figure 2). This prediction is what the speaker will try to communicate.

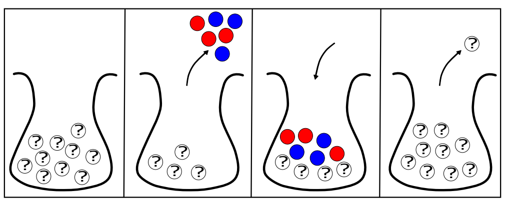

Figure 2: Partial observation of the content of the urn.

The probability of observing o red balls when the speaker draws a balls and there are s red balls in the urn is given by the hypergeometric distribution. Assuming that the agent has a prior belief distribution over the state space S, we can say that each pair $\langle o, a \rangle$ induces a posterior rational belief distribution over S, computed as the Bayes-inverted hypergeometric distribution:[2]

$$\text{rat.bel}(s|o,a) \propto \text{hypergeom}(o; a, s, 10) * \text{prior}(s) \qquad (1)$$

On the basis of the rational belief resulting from her observation, together with the lexical meaning of the available messages, the speaker chooses the best message to send given her communicative goal.

Messages and semantics The speaker sends messages of the form *It is* $[\dots]$ *that a randomly drawn ball will be red*, choosing from the following 12 expressions to fill the gap:

likely	possible	unlikely
certainly likely	certainly possible	certainly unlikely
probably likely	probably possible	probably unlikely
might be likely	might be possible	might be unlikely

[2]For convenience, the prior distribution over states is assumed to be a symmetric betabinomial distribution between 0 and 10 with shape parameters $\alpha = \beta$ free in the model.

Simple messages (*likely*, *possible*, *unlikely*) have a simple threshold semantics:

$$[\![\text{likely}(p)]\!] = \{s \in S \mid s/10 > \theta_{likely}\}$$
$$[\![\text{possible}(p)]\!] = \{s \in S \mid s/10 > \theta_{possible}\}$$
$$[\![\text{unlikely}(p)]\!] = \{s \in S \mid s/10 < 1 - \theta_{likely}\}$$

The thresholds θ_{likely} and $\theta_{possible}$ are free parameters in the model (more about this below). The variable p can be instantiated with a sentence such as *A randomly drawn ball will be red*. For example, this semantics states that the meaning of *It's possible that a randomly drawn ball will be red* is the set of states where the objective probability of the ball being red is bigger than a certain threshold $\theta_{possible}$.

The semantics of complex messages is stated in a general form as follows:

$$[\![\texttt{modifier}[\texttt{simple}](p)]\!] = \{\langle o, a \rangle \mid \sum_{s \in [\![\texttt{simple}(p)]\!]} rat.bel(s|o,a) > \theta_m\}$$

where θ_m is the semantic threshold associated with the modifier.[3] Each state in the meaning of the simple message $[\![\texttt{simple}(p)]\!]$ is associated with a certain probability mass according to the rational belief induced in the speaker by each pair $\langle o, a \rangle$; the meaning of the complex message is computed collecting the pairs $\langle o, a \rangle$ where the probability mass of the states in $[\![\texttt{simple}(p)]\!]$ is greater than the semantic threshold of the modifier. The semantics of complex messages is rooted in the literal semantics of the simple ones. The difference between the two is that while the meanings of simple messages contain states of affairs, the meanings of complex expressions contains pairs denoting partial observations, i.e. distributions over states. Still, both simple and complex expressions can be linked to sets of probability distributions over world states. Making use of this allows for a uniform grounding of semantic meaning in a model of rational communication.

Beliefs and expected utility On the basis of the literal meaning of each message, we compute their effect on the so-called "literal listener", a theoretical construct modeling the interpretation process of a non-pragmatic agent. Each simple message induces exactly one belief distribution in the literal listener, whereas each complex message induces a set of distributions (one for each pair $\langle o, a \rangle$ in the meaning of the expression). This idea is captured in Equation 2, where the set of distributions $lit.bel$ is defined by cases as a function of messages.[4]

[3]We assume $\theta_{probably} = \theta_{likely}$ and $\theta_{might} = \theta_{possible}$. The threshold of the remaining modifier $\theta_{certainly}$ is free in the model.

[4]The delta function $\delta_{s \in [\![m]\!]}$ gives 1 as output if the state s belongs to the meaning of s, 0 otherwise. The expression $\text{rat.bel}(.|o,a)$ refers to the belief distribution over states induced in a rational agent by the observation of o red balls out of a.

$$\text{lit.bel}(m) = $$
$$= \begin{cases} \{P \in \Delta(S) \mid \forall s \in S : P(s) \propto \delta_{s \in [\![m]\!]} * \text{prior}(s)\} & m \text{ simple} \\ \{P \in \Delta(S) \mid \exists \langle o,a \rangle \in [\![m]\!] : P = \text{rat.bel}(.|o,a)\} & m \text{ complex} \end{cases} \quad (2)$$

We assume that the communicative goal of the speaker is to maximize the information transferred to the listener. Here we formalize this concept as choosing the message which brings the listener's factual beliefs as close as possible to the speaker's, i.e. which minimizes the distance between the probability distributions expressing these beliefs. In general each message is associated with a set of probability distributions over states, according to Equation 2. Idealizing, we assume that the literal listener would uniformly sample from this set of uncertain beliefs upon hearing each message. For this reason the expected utility (EU) of a message m given an observation $\langle o,a \rangle$ is defined as the negative average Hellinger distance between the speaker's belief distribution given an observation and the set of the listener's distributions given a message (Equation 3).

$$\text{EU}(m;o,a) = -\text{avg}\left[HD\left(\{\text{rat.bel}(.|o,a)\}, \text{lit.bel}(m)\right)\right] \quad (3)$$

where HD denotes a function computing pairwise Hellinger distances between two sets of discrete distributions.[5]

Production and interpretation Adopting the terminology of rational choice theory, the speaker's behavior is to soft-maximize the EU of each message given her observation:

$$\text{speaker.prob}(m|o,a) \propto exp(\lambda * \text{EU}(m;o,a)) \quad (4)$$

EU is multiplied by a rationality parameter λ (free in the model) that modulates "how rational" the choice is.[6] The distribution over messages defined in Equation 4 gives rise to the first half of the set of predictions made by our model, whose fit to the experimental data is discussed below.

A pragmatic listeners reasons about the received message and her model of speaker's behavior in order to infer the most likely interpretation. The pragmatic listener's behavior is modeled as the joint Bayesian inference over the variables of interest:

$$\text{listener.prob}(s,o,a|m) \propto \text{speaker.prob}(m|o,a) * \text{priors} \quad (5)$$

We are interested in the interpretation of uncertainty expressions alongside two axes of their communicative effect. One is the objective state of affairs communicated (i.e., the inference of s). The other is the subjective, high-order, state of uncertainty of the speaker (i.e., the inference of $\langle o,a \rangle$). The joint distribution defined in Equation 5 gives rise to the second half of the set of predictions made by our model, whose fit to the experimental data is discussed below.

Experiments

We conducted two experimental studies, a production task and an interpretation task. The goal of the production task was to collect human data on the use of simple and complex uncertainty expressions under different high-order uncertainty conditions. The goal of the interpretation task was to collect human data on the interpretation of the expressions in terms of inference of s,o,a.

Participants 252 self reported English native speakers with USA IP-addresses were recruited via Amazon's Mechanical Turk. 102 participants completed the production task, 150 participants completed the interpretation task.

Material Participants in the production task were exposed to visual stimuli depicting partial observations of the urn. We asked participants to imagine drawing a number of balls (access) and counting the red balls among them (observation). Then they would put the balls back in the urn, and make a prediction about the color of another randomly drawn ball (Figure 2).

The experimental conditions are the different observation/access configurations displayed to the participants. We selected 15 such configuration:

high	0/2	1/4	2/4	3/4	2/2
low	0/8	2/8	4/8	6/8	8/8
none	2/10	3/10	5/10	7/10	8/10

Each fraction in the table corresponds to a possible partial observation, e.g. 3/4 means accessing 4 balls and observing that 3 of them are red. The fractions are grouped according to their level of high-order uncertainty. Access values smaller than 5 balls are labeled "high" high-order uncertainty, whereas values greater than 5 correspond to "low" high-order uncertainty, and values equal to 10 represent no high-order uncertainty whatsoever.

The set of stimuli for the interpretation task was derived from the 12 expressions assumed in the model.

Procedure Before the experimental phase, participants completed a training phase which contained a cover story introducing an interactive game between two players, a sender and a receiver. Participants in the production task were told that they would play as senders, and that other players would receive their messages and try to guess the content of the urn. Participants in the interpretation task were told that they would play as receivers. The motivation for this setup was to

[5]Goodman and Stuhlmüller (2013) use Kullback-Leibler divergence as a measure of discrepancy between speaker and listener beliefs. We found Hellinger distance a more adequate measure in the present setting because utilities in terms of KL-divergence lead to speakers who will never use messages that are semantically false, whereas HD allows messages to be send if they are "true enough." The Hellinger distance between two discrete distributions P and Q is defined as $\text{HD}(P,Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_i\left(\sqrt{P_i}-\sqrt{Q_i}\right)^2}$.

[6]As $\lambda \to \infty$, the choice approaches perfect rationality.

clarify the purpose of the conversation when producing uncertainty expressions and to prompt participants to reason about the effect of their choices on other agents.

Each participant in the production task completed 12 trials, one for each of 12 conditions randomly picked from the 15 total conditions. In each trial the participant made the partial observation of the urn corresponding to the selected condition and was asked to send a message containing a prediction about the color of a randomly drawn ball. Crucially, this prediction must be expressed by completing a sentence of the form *It* [...] [...] *that the next ball will be red*, selecting the most appropriate combination of auxiliary/modifier and uncertainty expression from two drop-down menus (Figure 3).[7]

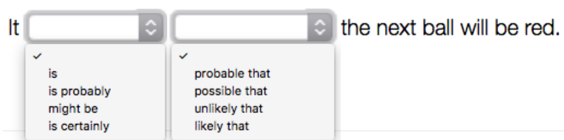

Figure 3: Input menus in the production task.

Each participant in the interpretation task completed 24 trials, 2 for each of the 12 expressions. That is, for each expression there were 2 kinds of trials, perfectly balanced, in random order. Half of the trials ("state" trials) recorded participants' interpretation of the message alongside the objective axis, i.e. their answer to the question "How many red balls do you think there are in the urn?", expressed with a natural number selected with a slider ranging from 0 to 10. Half of the trials ("observation" trials) recorded participants' interpretation of the message alongside the subjective uncertainty axis, i.e. their answers to the questions "How many balls do you think the sender has drawn? And how many of them do you think were red?", expressed with two natural numbers selected on sliders ranging from 0 to 10 (Figure 4).

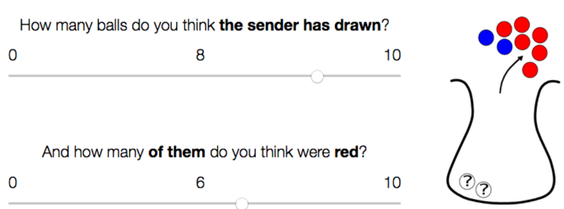

Figure 4: Input sliders in observation trials. The picture on the right dynamically visualized the current slider selection in order to provide immediate visual feedback for a selection.

Results Results are visualized in Figures 5 and 6 and will be discussed in the light of the model's predictions below.

[7]The possible choices included not only *likely* but also *probable* in embedded position. However, having not found interesting differences in the behavior of these two expressions, the results reported in this paper, the visualization in Figure 5 and the model evaluation are all based on data in which the counts of participants' choices of messages containing *probable* and *likely* have been aggregated.

Model evaluation and criticism

Model fit The data collected in the production task and the interpretation task are respectively counts of expression choices in each observation condition, and counts of state, access and observation choices for each expression. We used the data to compute credible values for the free parameters of the model, i.e. the semantic thresholds θ_{likely}, $\theta_{possible}$, $\theta_{certainly}$, the shape parameter of the prior belief distribution α, the rationality parameter λ. We implemented the computational model in JAGS (Plummer, 2003) and approximated the posterior distribution of parameters given the experimental data. We assumed flat prior distributions over the parameters with support $[0;1]$ for the semantic thresholds and $[0;20]$ for α and λ. We gathered two chains of 2500 samples after an initial burn-in of 2500. We checked convergence via $\hat{R}$ (Gelman & Rubin, 1992). Each sample consists of a vector of inferred values for each parameter. The following table summarizes the mean values for the threshold parameters together with their 95% highest density intervals (HDIs):[8]

	θ_{likely}	$\theta_{possible}$	$\theta_{certainly}$
mean	0.531	0.214	0.979
HDI	0.511-0.551	0.200-0.236	0.965-0.996

Notice that the model recovers plausible values for thresholds given the data without assuming them from the start.

For each sample vector of parameter values our model generates a set of predictions about speaker's and listener's behavior. In order to evaluate our model we correlated each set of predictions with the set of corresponding experimental count data. The results are collected in vectors of Pearson's correlation scores, whose means and HDIs give us an indication of the overall performance of the fitted model, as summarized in the following table:

	expression	*state*	*access*	*observation*
mean	0.649	0.862	0.883	0.941
HDI	0.647-0.651	0.857-0.867	0.880-0.886	0.938-0.943

Discussion Correlation scores do not provide detailed information about what aspects of the data the model can and cannot explain. To get a better sense of the performance of the model we compare data and predictions in more detail with posterior predictive checks (PPCs) (Kruschke, 2014).

We begin with the production task (Figure 5). Visual inspection of the plot suggests interesting features of the data. First, the number of observed red balls seems to have an influence on the choice of expressions. For example, with the same access of 8 (middle row of Figure 5), different observation values (0, 2, 4, 6 and 8) resulted in different distributions of expressions. This is an intuitive result, and the model correctly predicts the general pattern. Second, the same proportions of red balls but with different access levels seem to result in different expression choices. For example, compare the distributions of expressions observed (and predicted) with

[8]The other parameters of interest are α: mean= 6.373, HDI: $5.546-7.178$; and λ: mean= 5.429, HDI: $5.192-5.659$.

a proportion of 0 observed red balls and access values equal to 2 and 8, and similarly with a proportion of 1 and access values equal to 2 and 8. The distributions are different, and the model seems to predict the patterns.

However, there are also several discrepancies between observed data and the models PPCs. Discrepancies show in Figures 5 and 6 whenever the HDIs of the PPCs do not include the observed frequencies: in these cases the model, being trained on the data, would still be surprised, so to speak, by seeing the data points where observations do not fall in to the HDIs of our PPCs. For example, the model underpredicts choice frequencies of *might be possible* in favor of *possible* in the high uncertainty conditions and underpredicts *unlikely* and *likely* in the no uncertainty conditions. More in general, the model almost always overpredicts choice of, e.g., *certainly possible* and *might be unlikely*. At the same time, whenever PPCs are off for simple expressions, the model underpredicts their choice frequency. This suggests that a crucial ingredient might be missing from the model, namely a baseline preference of some expressions over others.

Turning now to the interpretation task (Figure 6), we observe that in general the patterns displayed in the data seem to be captured relatively well by the model. However, PPCs highlight a number of discrepancies. One clear example concerns the state interpretation for *unlikely* and its nested versions (left panel, right column): the predictions are visibly shifted to the right compared to the data. Another feature that the model fails to predict is the relatively low counts of access choices of 5 (compared to 4 and 6) for several expressions (middle panel), although this seems to be a puzzling feature of the data rather than an obvious shortcoming of the model.

Conclusion

Communication under high-order uncertainty raises a number of issues for formal semantics and pragmatics. Our work here is intended as a first but transparent explication of a number of assumptions that allow the formulation of a computational model of the use and interpretation of complex uncertainty expressions. The resulting model captures basic patterns in the data well enough, suggesting that our basic assumptions are not entirely off. Still, detailed model criticism also reveals a number of shortcomings. These point the way to further exploration; we see our main contribution exactly in this pointing. Most importantly, a measure of a differential inclination to produce messages (e.g., in terms of frequency, length, salience) should be included. Also, the artificial restriction on the set of message choices should ideally be relaxed as much as possible. Moreover, it will be telling to see how participants react to contextual manipulations such as of the relative relevance of information about the world state vs. information about the speaker's epistemic state.

Acknowledgments

Financial support by the Institutional Strategy of the University of Tübingen (Deutsche Forschungsgemeinschaft, ZUK 63) is gratefully acknowledged.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1991). Is human cognition adaptive? *Behavioral and Brain Sciences*, *14*(03), 471–485.

Beyth-Marom, R. (1982). How probable is probable? a numerical translation of verbal probability expressions. *Journal of Forecasting*, *1*(3), 257–269.

Egan, A., & Weatherson, B. (2011). *Epistemic modality*. Oxford University Press.

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

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, *7*, 457–472.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, *5*(1), 173–184.

Herbstritt, M., & Franke, M. (2016). Definitely maybe and possibly even probably: efficient communication of higher-order uncertainty. In A. Papafragou, D. Grodner, D. Mirman, & J. Trueswell (Eds.), *Proceedings of the 38th annual conference of the cognitive science society.*

Kratzer, A. (1991). Modality. In A. von Stechow & D. Wunderlich (Eds.), *Semantics: An international handbook of contemporary research* (pp. 639–650). de Gruyter.

Kruschke, J. (2014). *Doing Bayesian Data Analysis, 2nd Edition: A tutorial with R, JAGS, and Stan*. Academic Press.

Lassiter, D. (2011). *Measurement and modality: the scalar basis of modal semantics*. Unpublished doctoral dissertation, NYU Linguistics.

Moss, S. (2015). On the semantics and pragmatics of epistemic vocabulary. *Semantics and Pragmatics*, *8*(5), 1–81.

Plummer, M. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of DSC3* (Vol. 124, p. 125).

Swanson, E. (2006). *Interactions with context*. Unpublished doctoral dissertation, Massachusetts Institute of Technology, Cambridge MA.

Teigen, K. H. (1988). When are low-probability events judged to be 'probable'? effects of outcome-set characteristics on verbal probability estimates. *Acta Psychologica*, *6*(2), 157–174.

Windschitl, P. D., & Wells, G. L. (1996). Measuring psychological uncertainty: Verbal versus numeric methods. *Journal of Experimental Psychology: Applied*, *2*(4), 343.

Windschitl, P. D., & Wells, G. L. (1998). The alternative-outcomes effect. *Journal of Personality and Social Psychology*, *75*(6), 1411–1423.

Yalcin, S. (2010). Probability operators. *Philosophy Compass*, *5*(11), 916–37.

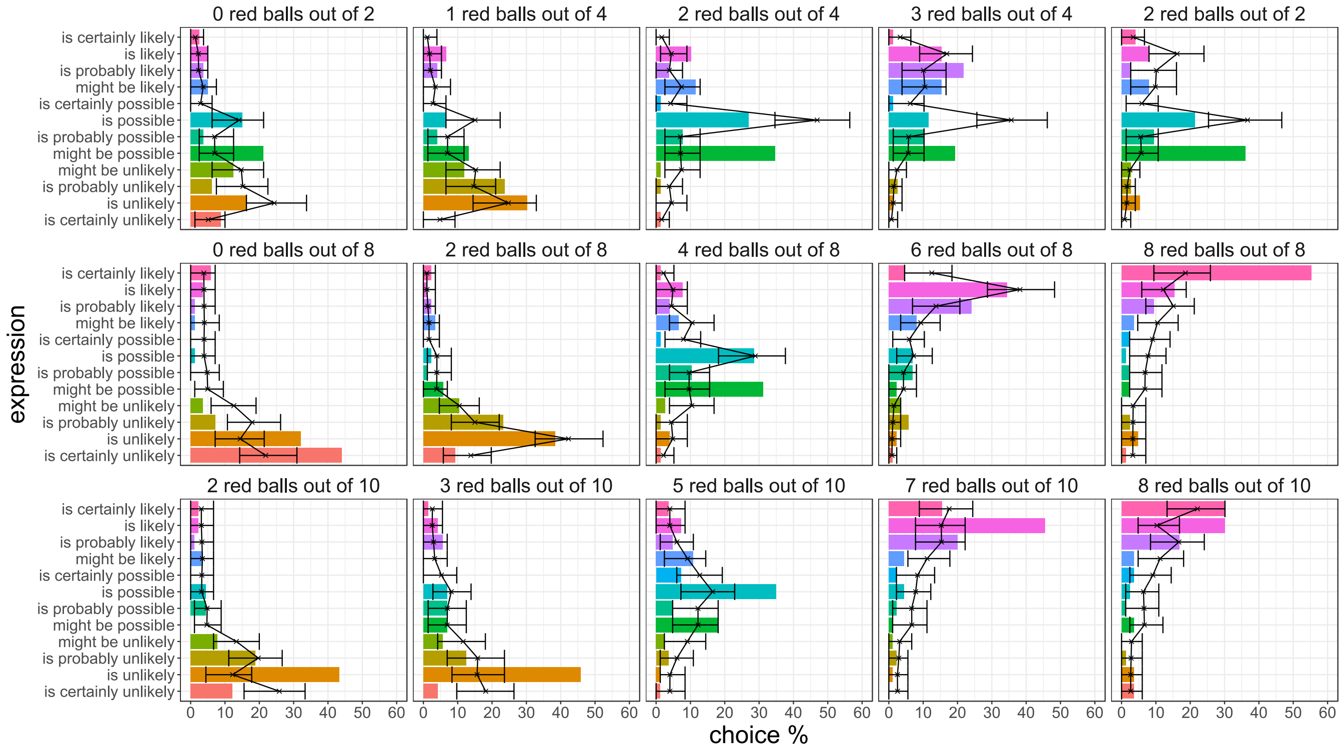

Figure 5: Percentages of expression choices in each partial observation condition, together with mean predictions and HDIs.

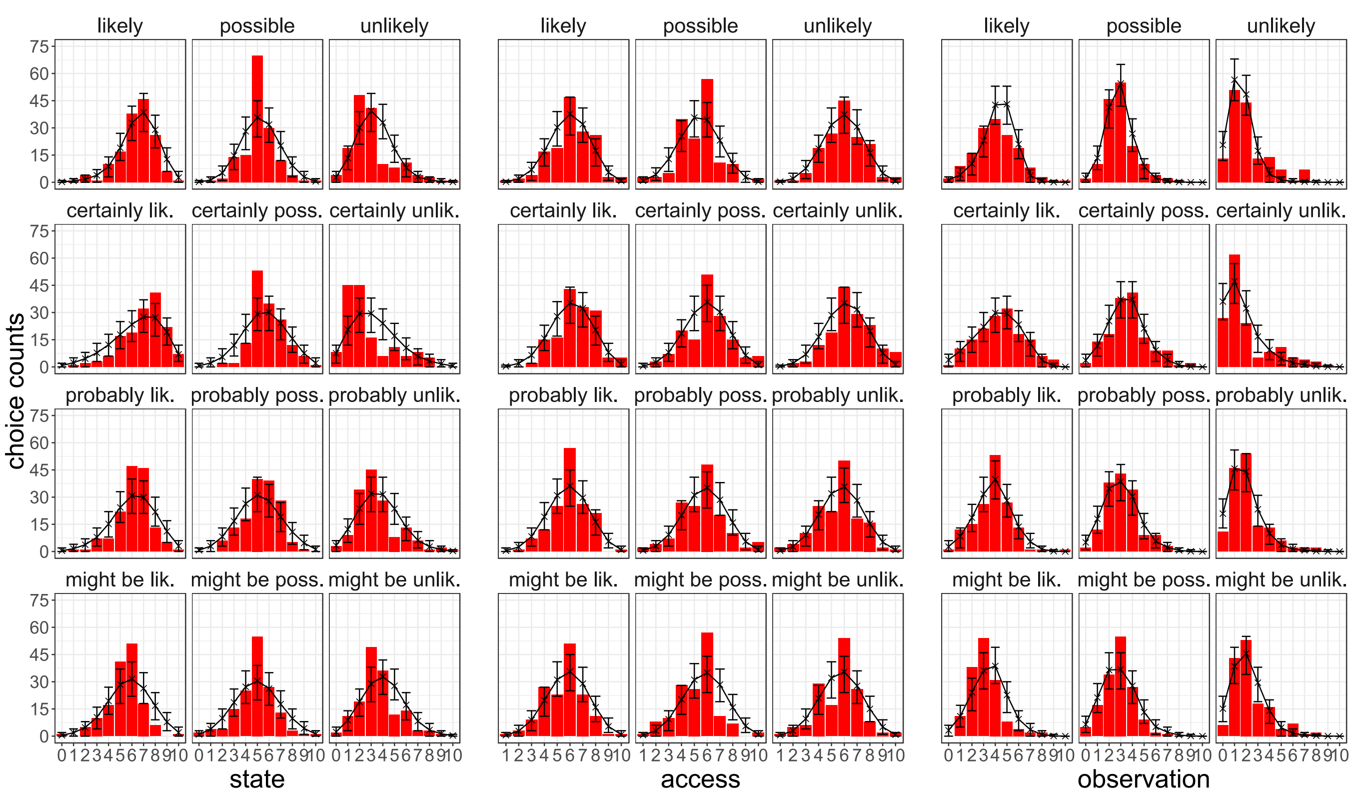

Figure 6: Counts of state, access and observation choices for each expression, together with mean predictions and HDIs.