Towards a model of confidence judgments in concept learning

Tracey E. Mills*, Tony Chen*, Cedegao E. Zhang* & Joshua B. Tenenbaum
{temills, thc, cedzhang, jbt}@mit.edu
Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology, Cambridge, MA 02139, USA
*These authors contributed equally to this work.

Abstract

Confidence is an important concept in cognitive science, as it integrates seamlessly with our beliefs, goals, and decisions. Humans naturally represent and express degrees of confidence in beliefs and predictions that reflect their accuracy. However, the dynamics of how our underlying beliefs about the world relate to explicitly represented confidence over those beliefs is yet not well understood. In this work, we make progress on this question in the domain of *concept learning*. Specifically, we analyze how confidence and beliefs jointly evolve in the absence of explicit feedback. We evaluate some leading computational accounts of confidence in the present literature, and we find that these accounts do not accurately predict confidence in the context of our task. We advocate for caution in making claims about the generalizability of such accounts across tasks or domains and propose some new model-based measurements for predicting humans' confidence judgments. Of these measurements, we find that ones taking individual-level response patterns into account perform the best. We close by suggesting promising future directions for the study of confidence in concept learning.

Keywords: confidence judgments; subjective belief; concept learning; language of thought; Bayesian inference

Introduction

Humans display a remarkable ability to accurately quantify their confidence in their beliefs and predictions and to act accordingly. When faced with an uncertain situation, a person may choose to wait and gather more information, take a chance on a risky bet, or give up on a confusing problem. More generally, having a calibrated sense of confidence helps guide our actions. We do not cross the street if we think a car might be about to round the corner, unless there is a better chance that the noise behind us came from a charging bear. Consideration of one's own degree of belief thus supports robust and flexible decision making and planning, and is key to how we represent the world and our place in it.

Our language reflects this fact, with confidence expressible in a myriad of ways. Sampling from English, "I think that...", "likely", "might", "maybe", and "hope" each convey subtly different flavors of the graded nature of belief. Indeed, although different linguistic expressions of graded belief have often been treated as indicators of some common phenomenon in past work (i.e. *confidence*, *certainty*, and *belief* used interchangeably on measurement scales), the validity of this assumption requires verification (cf. Pouget, Drugowitsch, & Kepecs, 2016).

Though humans naturally refer to and represent their beliefs as graded and rely on graded belief representations to act intelligently in the world, it is not clear how these representations are formed, and how those representations are translated into explicit numerical judgments of confidence. Here, we investigate hypotheses for how one's representation of their degree of belief, or *confidence* as we refer to in the remainder of this paper, might be computed during learning. Specifically, we study human judgments of confidence in a sequential concept learning paradigm in which participants are asked both to make predictions and rate their degree of belief in their prediction being correct. We analyze how people's beliefs and confidence jointly evolve as they gather additional evidence, despite them never receiving explicit feedback on the correctness of their predictions. We compare the predictive accuracy of a battery of behavioral and model-based measures on these confidence ratings across various word conditions (i.e. *confidence*, *certainty*, and *belief strength*). Notably, we evaluate several measures proposed to account for confidence in perceptual discrimination domains on our task to evaluate the extent to which those leading accounts of confidence might be domain generalizable.

In the remainder of this paper, we will first establish a positive result regarding the exchangeability of the words used in each condition, finding no meaningful differences in the effects of the word used to prompt participants for confidence judgements. We then show that although model-based measures computed from a Bayesian posterior over beliefs demonstrate significant predictive validity, the best predictor of participant confidence ratings is one that is in principle not computable given the lack of feedback in our experiment: the correctness of a participant's predictions. Finally, we discuss promising approaches for closing the gap between what people seem to know about their own belief validity, and theories of how this knowledge might be computed.

Related works

In research on linguistic semantics, recent years have witnessed a line of work devoted to the study of the meaning of words related to (un)certainty such as "confident" (Cariani, Santorio, & Wellwood, 2022), "certainly" (Lassiter, 2017), and "believe"(Hawthorne, Rothschild, & Spectre, 2016). Traditionally, philosophers and linguists regard belief as a binary concept—one either believes some proposition *P* or not (Hintikka, 1962; von Fintel & Heim, 2011). However, semanticists have recently found graded semantic representations of

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

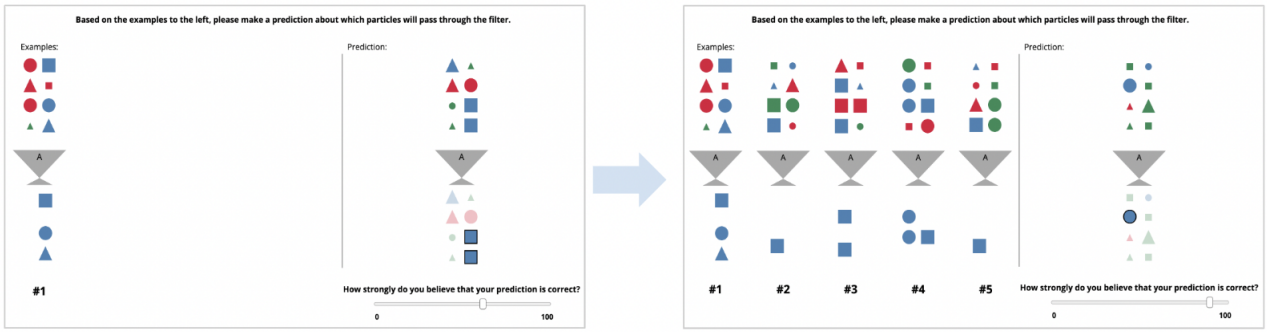


Figure 1: The experiment setup as seen by participants. Arrows indicate the flow of the experiment in a simple block. Participants see 6 blocks of single filter trials as above, where they were presented with 5 examples of an unknown filter acting on a set of inputs, and asked to make predictions on a novel set of particles, along with providing their confidence in the prediction. Observations and predictions were interleaved such that participants would observe one example, then predict on another set, and so on. All previously seen examples were shown to the participant on the left side to control for any effects of working memory maintenance.

belief words generally more appealing based on linguistic evidence (Goodman & Lassiter, 2015; Hawthorne et al., 2016), which converges to the long-held assumptions in Bayesian cognitive science.

Past work aiming to characterize how confidence and related concepts are computed in the mind has focused largely on perceptual discrimination tasks in which participants must determine their confidence in a forced choice between two or a few options (Maniscalco & Lau, 2012; Mamassian, 2016; Adler & Ma, 2018; Rahnev, 2020). People's judgments in these tasks typically closely track the probability of a correct response under a signal detection framework, leading some to theorize that confidence in general arises from the computed *probability* of some prediction being correct (Drugowitsch, Moreno-Bote, & Pouget, 2014). And deviations from this pattern can be explained by appealing to metacognitive failures such as the incorrect weighting of sensory signals (Shekhar & Rahnev, 2021).

While this line of work establishes a relationship between posterior probabilities and confidence, it is unclear how these results might generalize to different tasks or domains. Li and Ma (2020) find that even adding one additional option to a two-alternative forced choice perceptual discrimination task raises complications for this story. In this case, average participant confidence was not best predicted by the probability of the correct response, but rather by the *difference* in probability between the best two responses. This finding importantly suggests that confidence does not map simply onto singular posterior probabilities in some distribution, but rather involves non-local statistics from this distribution as a whole.

However, characterizing this involvement becomes more complicated in tasks or domains in which the set of choices is extremely large or even potentially infinite. For example, many concept learning tasks require the formation and evaluation of beliefs about the meaning of some concept where the space of possible meanings is unbounded, defined in terms of abstract compositional rules. While it is still natural to express degrees of confidence over beliefs in such tasks, it is less obvious how these might be computed. It has been suggested that people might approximate the probability of certain beliefs through their sampling propensity from a probability distribution implicitly defined in terms of a generative model for the space of alternatives (Icard, 2016; Vul & Pashler, 2008; Vul, Goodman, Griffiths, & Tenenbaum, 2014). Concept learning tasks thus provide an exciting test-bed for studying whether and how confidence might correspond to probability distributions over beliefs, even when these distributions are not fully accessible.

Martí, Mollica, Piantadosi, and Kidd (2018) makes progress on this question by considering how a person's certainty that they know the meaning of a Boolean concept might map onto the true probability distribution over possible concept meanings. Participants in this task repeatedly make predictions about whether an object is described by the concept, as well as whether or not they are certain that they know what the concept is. They are then immediately told whether or not their prediction was correct. Martí et al. (2018) find that while people's confidence has a strong relationship with their performance on the task, extrinsic cues such as recent feedback better predict participant confidence than attributes of the normative probability distribution over beliefs, such as its entropy or the probability of the most probable belief. This result is interpreted as evidence that people derive confidence judgements primarily from external sources of information about how confident they *should* be rather than from some attribute of the belief representation itself.

Approach

In many real-world instances of concept learning and belief formation, people do not receive such explicit feedback. Indeed, Martí et al. (2018) find that even during one shot learning, people express degrees of belief that track their success. It seems unlikely that confidence is then fully determined by

external cues.

How might this puzzle be resolved? Here, we provide initial evidence that part of an answer lies in considering how confidence might be derived from representations of probability which are plausibly accessible to a particular individual, perhaps derived from approximations of some normative probability distribution over beliefs.

We do so by assessing participants confidence in an iterative Boolean concept learning paradigm with a large space of possible concept meanings, and in which no direct feedback is provided. Instead, participants alternate between accruing additional evidence about a concept, and making predictions and judgements about their confidence in their predictions. Thus, participants might only assess their own performance by internally monitoring and updating the probability of their predictions as they incorporate additional evidence. We then examine the relationship between reported confidence and both previously established global metrics, as well as individual-specific metrics, derived from a Bayesian language of thought concept learning model.

Experiment

We employ a Boolean concept learning task (Feldman, 2000; Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Piantadosi, Tenenbaum, & Goodman, 2016), in which participants learn and make predictions about **filters** which allow only certain types of **particles** to pass through them. Filters act on groups of 8 particles at a time, and allow particles through according to key feature axes: size, shape, color, position in the group, and their logical combinations. For example, one filter might only let through particles that are red and triangle shaped, while another might only let through particles that are on the left side of the group or that are small. See Figure. 1 for a depiction of a filter that a participant might learn about.

Participants learn about filters through examples, make predictions, and rate their confidence in their predictions. We also examine how differently worded instructions might affect the way people report degrees of belief. Specifically, we ask people to report either how confident they are that they are correct, how certain they are that they are correct, or how strongly they believe that they are correct. While we do not expect participant behavior to vary significantly in the three conditions, we consider it important and theoretically interesting to test the assumption often made in the literature that these words tend to indicate the same underlying concept.

Procedure

We recruited participants from Prolific (N= 150) to take part in this task. Participants are evenly split between the three word conditions and see the same wording each time they are asked to give a rating.

Participants sequentially learn about 6 different filters, randomly chosen from a group of 30 filter concepts. In a single filter block, a participant is first presented with an example of that filter acting on input particles, where the 8 input particles are seen above the filter and only the particles that are allowed through are seen below the filter. With this example still in view, participants then make a prediction about which of the particles in a new set of 8 will pass through the filter and rate their prediction according to their assigned word condition using a sliding scale from 0 to 100. Participants repeat this alternating process of example and prediction until they have seen 5 examples and given 5 predictions and ratings for that filter. All examples seen so far remain on the screen when making predictions, ensuring that participants can always access and remember their previous examples.[1]

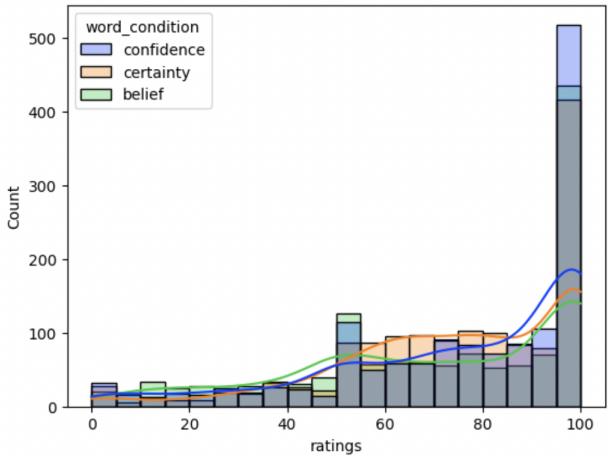


Figure 2: Histograms of ratings in all trials in the three word conditions.

Behavioral Results

People's judgments in each of the word condition groups are shown in Figure 2. We observe substantial overlap between each of the three word conditions. Participants' judgements across word conditions exhibit reasonably high correlations ($r = 0.75$ for "confidence" vs "certainty", $r = 0.70$ for "confidence" vs "belief", and $r = 0.62$ for "certainty vs belief", $p < .0001$ for each). For a point of reference, we compute the split-half reliability within word conditions. We find that on average for the between group correlations, $r = 0.73$, while for the within-group correlations, $r = 0.72$, across 20 such random groupings. These results tentatively support the putative semantic similarities of the three words. Thus, we pool participants together across word conditions for all subsequent analyses.

We find that across concepts, participants' reported confidence in their predictions has a strong positive relationship with their accuracy ($r = 0.713$, $p < .0001$), suggesting that participants' computed confidence is well-calibrated. We also

[1] After every two filter blocks, we also asked participants to give 3 additional predictions and ratings, interleaved with two examples, of the output of the two previously learned filters stacked on top of one another. Participants are told that in such trials, the particles will first be sent through the top filter, and the particles which pass through that filter will then be sent through the bottom filter. We do not analyze these compositional filter blocks in this work.

Table 1: The context-free grammar used to generate hypotheses.

Nonterminal		Expansion	Description
START	$\rightarrow$	$\lambda\, D.$ LIST	A function that takes D (input data) and returns a list of 1s and 0s
LIST	$\rightarrow$	*everything*	A list representing that every particle goes through
	$\rightarrow$	*nothing*	A list representing that particle goes through
	$\rightarrow$	($\wedge$ LIST LIST)	Each entry is 1 if it is 1 in both argument lists
	$\rightarrow$	($\vee$ LIST LIST)	Each entry is 1 if it is 1 in either argument lists
	$\rightarrow$	($\neg$ LIST)	Each entry is $1 - x$ for x in the argument list
	$\rightarrow$	(f D)	f applies to D, which returns a list
f	$\rightarrow$	COLOR	Color features
	$\rightarrow$	SHAPE	Shape features
	$\rightarrow$	SIZE	Size features
	$\rightarrow$	POSITION	Position features
COLOR	$\rightarrow$	*red?*	(For each entry) 1 if the particle is red else 0
	$\rightarrow$	*green?*	1 if the particle is green else 0
	$\rightarrow$	*blue?*	1 if the particle is blue else 0
SHAPE	$\rightarrow$	*circle?*	1 if the particle is a circle else 0
	$\rightarrow$	*square?*	1 if the particle is a square else 0
	$\rightarrow$	*triangle?*	1 if the particle is a triangle else 0
SIZE	$\rightarrow$	*big?*	1 if the particle is big else 0
	$\rightarrow$	*small?*	1 if the particle is small else 0
POSITION	$\rightarrow$	*top?*	1 if the particle is among the top four else 0
	$\rightarrow$	*bottom?*	1 if the particle is among the bottom four else 0
	$\rightarrow$	*left?*	1 if the particle is among the left four else 0
	$\rightarrow$	*right?*	1 if the particle is among the right four else 0

find that a participant's average accuracy across the current and past predictions for a concept is a very slightly numerically better predictor of confidence than accuracy on the current prediction alone ($r = 0.714$, $p < .0001$). This result is consistent with the claims of Martí et al. (2018) that prior success adds power in predicting current confidence. Importantly, we find that this is the case even when participants do not receive feedback about their past success.

Modeling

We take as a starting point standard language of thought (LoT) models, which model concept learning as program induction (Piantadosi, 2011). Specifically, filter concepts are defined to be programs generated by a probabilistic context-free grammar (PCFG), and learning a filter concept involves identifying the programs that best explain the observed filter behavior. Motivated by the idea that people determine plausible latent world states based on observed data in accordance with principles of Bayesian statistics (Tenenbaum, Kemp, Griffiths, & Goodman, 2011), our model does so by approximating the posterior probability distribution over programs.

Our PCFG, depicted in Table 1, implements a prior distribution over filter concept programs. These programs take a list of particles as input and return a list of 0s and 1s indicating whether each input particle "passed through" (1) or not (0). Programs are made up of primitive concepts: color concepts (e.g. *red*, determining that only red particles pass through), size concepts (e.g. *big*), shape concepts (e.g. *triangle*), position concepts (e.g *left*), and their logical conjunctions (e.g. *red AND triangle*). Additionally, we have trivial terminals 'everything' and 'nothing' that allow all or none of the particles to pass through, respectively.

The likelihood of an observation is calculated independently for each particle. The probability of a particle passing through is 1 when the outcome for particle i mirrors the program output and 0 otherwise.

The model learns a distribution over programs for a filter concept given observed input/output examples. Inference is performed using Sequential Monte Carlo with tree-rejuvenation metropolis hasting steps, to mirror the sequential structure of the task.

At each step, we compute the posterior predictive distribution, obtained by marginalizing over candidate hypotheses, weighted by the posterior probability of each hypothesis. Because participants rated confidence in their predictions and not about their hypotheses for a filter, all model metrics of confidence were calculated using this distribution, instead of the posterior distribution over hypotheses.

Figure 3: Learning curves of human participants and the Bayesian model.

Modeling results

Our model learns most concepts quickly and more effectively than humans. This indicates that while in many cases the correct belief can be correctly deduced by an ideal learner, humans likely have a more approximate representation of the probability distribution over beliefs which nonetheless supports the eventual learning of the concept in most cases.

For our analyses we consider the probability distribution produced by the model as a baseline for what might be learned in principle based on the evidence available. We later discuss methods humans might employ for approximating this distribution in a more resource-constrained manner.

We compute and compare multiple model-derived metrics based on how well they can predict participant confidence. These metrics are all transformations and statistics calculated from the posterior predictive distribution over outputs, obtained by marginalizing over all possible hypotheses.

We first calculate the **max probability** metric, which is simply the probability of the best output under the posterior predictive distribution. We also evaluate several other measures that take into account the probability of alternatives. These are the **difference** and **ratio** measures proposed by Li and Ma (2020), computed by taking the difference and the ratio of the probabilities of the two best outputs under the posterior predictive distribution, respectively, as well as the **entropy** of the posterior predictive distribution, which Martí et al. (2018) found to be one of the most predictive model metrics in their concept learning paradigm.

While the metrics discussed above make the same predictions for each individual, we might also consider model metrics which make different predictions based on an individual's responses. Here, we consider **response probability**, which is the probability the model assigns to the particular response given by a participant. We also consider the **average response probability**, which is simply the response probability for an individual average across the current and previous predictions for the current filter. We find that response probability performs best out of the model metrics, with comparable predictive power as other behavioral measures. Response probability and response correctness are closely related, since our model generally assigns high probability to correct responses. However, the response probability explains additional variance over and above response correctness (a decrease in AIC to 1052 from 1062, and an increase in R^2 from 0.51 to 0.55), which may be mildly suggestive that the predictive power of response probability is not entirely due to it's correlation with response correctness.

Table 2: Confidence predictor results. The p-values for all the Rs are below 0.001.

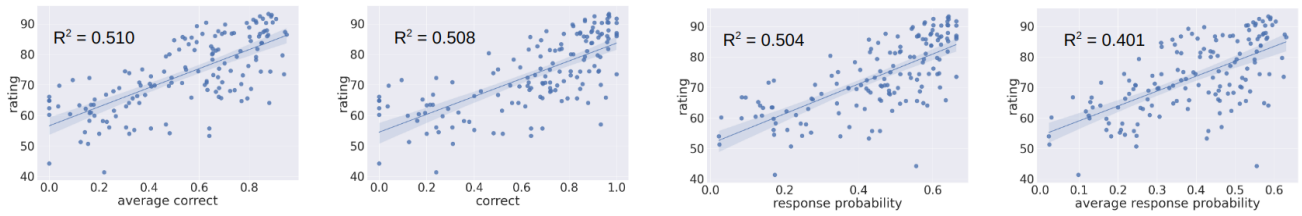
Predictor	R	R^2
Avg Correct	0.714	.510
Correct	0.713	.508
Response Probability	0.709	.504
Avg. Response Probability	0.633	.401
Difference	0.296	.088
Max Probability	0.287	.083
Entropy	-0.286	.082
Ratio	0.272	.074

Discussion

We find that leading hypotheses for confidence during perceptual discrimination, namely **max probability**, **difference**, and **ratio**, performed the worst out of the model metrics we tested. This suggests that we should be cautious when making general claims about how confidence judgments are related to a probability distribution over hypotheses, as this relation may in part depend on the task or domain. A both challenging and exciting step for future work is to evaluate the extent to which this disparate set of results might be united under a domain general account of confidence.

Global vs. individualized predictors

Past work has primarily focused on predicting human confidence through model metrics that do not take individual response patterns into account. Here, we find that there is indeed a strong relationship between confidence and the probability assigned by the model to the *actual response* given ($R = 0.709$), rather than the probability assigned to the best response. Because of the accuracy of our model, this metric approximates the response correctness metric in many cases. However, we find that it is not the case that the predictiveness of response probability is fully explained by its relationship with response accuracy. Additionally, unlike the correctness metric, which requires access to the ground truth, some approximation of the response probability is plausibly *accessi-*



(a) (b)

Figure 4: Regression plots for the *average correct*, *correct*, *response probability*, and *average response probability* predictors.

ble to humans as they make their prediction.

Incorporation of past performance

We find that the average accuracy across past and current trials performs almost identically with accuracy on the current trial when predicting confidence, despite not being perfectly correlated with each other. Therefore, while our best model-based metric predicts confidence based only on the posterior predictive distribution on the current trial, future work may additionally attempt to predict confidence based on additional factors that characterizes how one's probability distribution *changes* throughout learning.

At the same time, our finding that confidence is equally well predicted by a metric which does not take past performance into account stands in contrast to Martí et al. (2018), whose results instead suggest that confidence is primarily determined by past performance. These discrepancies may be explained in part by various task differences. Participants in our task did not have access to direct feedback on their predictions. When this external information is available, it may indeed play an important role in determining confidence. However, it is certainly not necessary for computing confidence, and real-world scenarios in which explicit feedback is available are likely to be the exception rather than the rule. Additionally, in our experiment concepts interact with 8 items at a time rather than 1, posing a greater computational challenge and potentially making it more difficult to understand and incorporate information from past judgements.

Future directions

In this work, we find that participants' confidence is well-calibrated to correctness, despite the complexity of our paradigm and the lack of explicit feedback. In fact, we find that no model metric predicts confidence as well as current and past accuracy does. In other words, the relationship between confidence and correctness is stronger than the relationship between confidence and the probability of correctness as predicted by our model. This suggests that people's confidence judgements encode some information about the correctness of their response which our model does not capture. We suggest two approaches for future work in light of this result.

Firstly, the explanatory gap between response probability and confidence judgments may be closed in part by models that form more human-like belief representations. Indeed, one limitation of this current work is the mismatch between the human and model learning trajectories. More resource-rational approaches might compute posterior representations based on one or a few posterior samples. These sparser representations might be more aligned with human distributions over hypotheses and the predictions that they make. In particular, certain model metrics such as posterior entropy require integrating over the entire support of the posterior, a computation that is intractable for most real-world problems. A more plausible account could involve parsimonious consideration of alternatives, instead of enumerating every single option. Other modifications might include probabilistic forgetting of prior evidence when computing the likelihood or ignoring of certain primitive types (e.g. color) during search.

Another important direction for future work is integrating models of metacognition with confidence judgments. It is possible that humans do not have direct access to the computations that support concept learning, nor a resulting posterior distribution over hypotheses. In that case, explicit confidence judgments may in part arise from higher-order computation: predictions are made by a cognitively impenetrable decision model, and an outer metacognitive model approximates the probability that the first-order decision maker is correct (Fleming & Daw, 2017). This higher-order model could incorporate external cues such as past feedback or generate its own feedback signal based on the prediction made.

Conclusion

Much of cognitive science holds that people's predictions and actions arise from a graded, implicit degree of belief. However, it is now well-established that this implicit measure does not map cleanly onto explicit judgments of confidence. We present preliminary evidence showcasing a gap between behavioral predictors of confidence and current computational accounts. We take a first step towards addressing this conceptual gap, by proposing other individual-level model-based measures, and advocating for a resource-rational perspective on confidence judgments. In the future, we hope to characterize at an algorithmic level the connection between the posterior in the head, and the confidence that people report.

Acknowledgements

We thank Sam Tenka for inspiring us to use the concept of filters for designing the experiment.

References

Adler, W. T., & Ma, W. J. (2018). Comparing bayesian and non-bayesian accounts of human confidence reports. *PLOS Computational Biology*, *14*(11), e1006572.

Cariani, F., Santorio, P., & Wellwood, A. (2022). *Confidence reports.* (Unpublished manuscript)

Drugowitsch, J., Moreno-Bote, R., & Pouget, A. (2014). Relation between belief and performance in perceptual decision making. *PloS one*, *9*(5), e96511.

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, *407*(6804), 630–633.

von Fintel, K., & Heim, I. (2011). Intensional semantics. *Unpublished lecture notes*.

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general bayesian framework for metacognitive computation. *Psychological review*, *124*(1), 91.

Goodman, N. D., & Lassiter, D. (2015). Probabilistic semantics and pragmatics: Uncertainty in language and thought. *The handbook of contemporary semantic theory, 2nd edition. Wiley-Blackwell*.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive science*, *32*(1), 108–154.

Hawthorne, J., Rothschild, D., & Spectre, L. (2016). Belief is weak. *Philosophical Studies*, *173*(5), 1393–1404.

Hintikka, J. (1962). *Knowledge and belief: An introduction to the logic of the two notions*. Cornell University Press.

Icard, T. (2016). Subjective probability as sampling propensity. *Review of Philosophy and Psychology*, *7*(4), 863–903.

Lassiter, D. (2017). *Graded modality: Qualitative and quantitative perspectives*. Oxford University Press.

Li, H.-H., & Ma, W. J. (2020). Confidence reports in decision-making with multiple alternatives violate the bayesian confidence hypothesis. *Nature Communications*, *11*(1), 2004.

Mamassian, P. (2016). Visual confidence. *Annual Review of Vision Science*, *2*, 459–481.

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and cognition*, *21*(1), 422–430.

Martí, L., Mollica, F., Piantadosi, S., & Kidd, C. (2018). Certainty is primarily determined by past performance during concept learning. *Open Mind*, *2*(2), 47–60.

Piantadosi, S. T. (2011). *Learning and the language of thought* (Doctoral dissertation). MIT.

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2016). The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological Review*, *123*(4), 392–424.

Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: distinct probabilistic quantities for different goals. *Nature neuroscience*, *19*(3), 366–374.

Rahnev, D. (2020). Confidence in the real world. *Trends in Cognitive Sciences*, *24*(8), 590–591.

Shekhar, M., & Rahnev, D. (2021). Sources of metacognitive inefficiency. *Trends in Cognitive Sciences*, *25*(1), 12–23.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive science*, *38*(4), 599–637.

Vul, E., & Pashler, H. (2008). Measuring the Crowd Within: Probabilistic Representations Within Individuals. *Psychological Science*, *19*(7), 645–647.