The logic of guesses: how people communicate probabilistic information

Tadeg Quillien, Christopher G. Lucas
School of Informatics, University of Edinburgh

Abstract

How do people respond to a question when they are not certain of the answer? Probabilistic theories of cognition assume that the mind represents probability distributions over possible answers, but in practice people rarely recite these probability distributions out loud: instead they make simple guesses. Consider how you would express your belief about how many people live in the European Union. You would probably not say "a Gaussian with mean 300 million and standard deviation 50 million" – you would make a simple guess, such as "between 200 and 400 million". Here we present a simple rational analysis of these guesses. We assume that communicating the full probability distribution in one's head would take too much time, so people offer simple guesses in order to communicate a compressed version of this distribution. Drawing on information theory, we show that it is possible to measure how well a guess encodes a given probability distribution, and suggest that people tend to make guesses that provide the best such encoding. Two experiments provide preliminary evidence for the model. Our theory explains from first principles why guesses seem to strike a balance between accuracy and informativeness.

Keywords: computational modeling; guesses; probability; information theory; judgment under uncertainty

Introduction

Your friend is about to reach inside the box depicted in Figure 1, and will randomly draw one ball while blindfolded. Which ball is going to come out? For a Bayesian, the answer is easy: the ball will be red with probability 5/12, green with probability 3/12, yellow with probability 3/12, and blue with probability 1/12. According to a prominent approach to cognition, the mind is approximately Bayesian (Tenenbaum et al., 2011; Oaksford & Chater, 2007): in this kind of simple problem, it should hold a representation of the probability distribution over possible outcomes. Yet, most people would not recite the whole probability distribution out loud if they were asked which ball will come out. Instead, they might say "it will probably be a red ball", or "it will be a red, green, or yellow ball". How do people make these simple guesses?

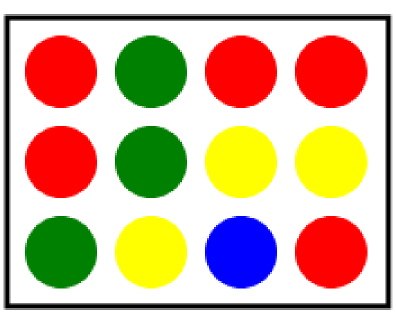

Figure 1: A box with colored balls.

Epistemologists have pointed out that guesses seem to have their own particular phenomenology (Holguin, 2022; Dorst & Mandelkern, 2021). Intuitively, guesses that are likely to come true tend to be better. For example "it will be a red ball" is a better guess than "it will be a blue ball". But the quality of a guess is not entirely determined by its probability: for example "it won't be a red ball" is more probable than "it will be a red ball", but the latter seems like a more natural guess (Dorst & Mandelkern, 2021). A challenge for a cognitive theory of guesses is to explain their peculiar logic.

A rational theory of guesses

We suggest that guesses (like "it will be a red, green or yellow ball") solve a compression problem. When people answer a question about an uncertain outcome, it would take too much time for them to explicitly recite the relevant probability distribution. Instead, they make a guess, which functions as a compressed representation of the probability distribution in their mind. The guess leaves out some information, but is nonetheless useful for most purposes. If you tell someone that the ball coming out of the box will be red, green, or yellow, they will not be able to infer the exact content of the box, but they may get a good enough approximation.

Thus, a guess is good to the extent that people hearing the guess can reconstruct the probability distribution over outcomes that the speaker had in mind. If you tell someone that the ball will be "red, green, or yellow", she can infer that red, yellow, and green are more probable outcomes than blue, but she has no reason to think that any of the three colors (red, yellow, green) is more likely than the others[1]. So, her best bet is to construct a probability distribution that looks like the one in Figure 2c. As another example, if you tell her "it will be a red ball", her best bet is to infer a probability distribution over outcomes that looks like the one in figure 2b.

More formally, guesses can be seen as implicitly encoding a probability distribution where all outcomes mentioned in the guess have equal probability, and have higher probability than outcomes not mentioned in the guess. To construct such a distribution, one can define the probability of each outcome x as:

$$P(X = x) = \delta(x, g) * \frac{1-\lambda}{n_g} + \frac{\lambda}{n}$$

where $\delta(x, g)$ is 1 if x is included in the guess, and 0 otherwise. n_g is the size of the guess (the number of possible out-

[1] We assume that the audience knows that there are 12 balls in the box, that they can be red, yellow, green and blue, but does not know in which proportions these colors are represented. Note that for simplicity here we consider settings where the audience knows what outcomes are possible, but our approach is compatible with situations where that is not the case.

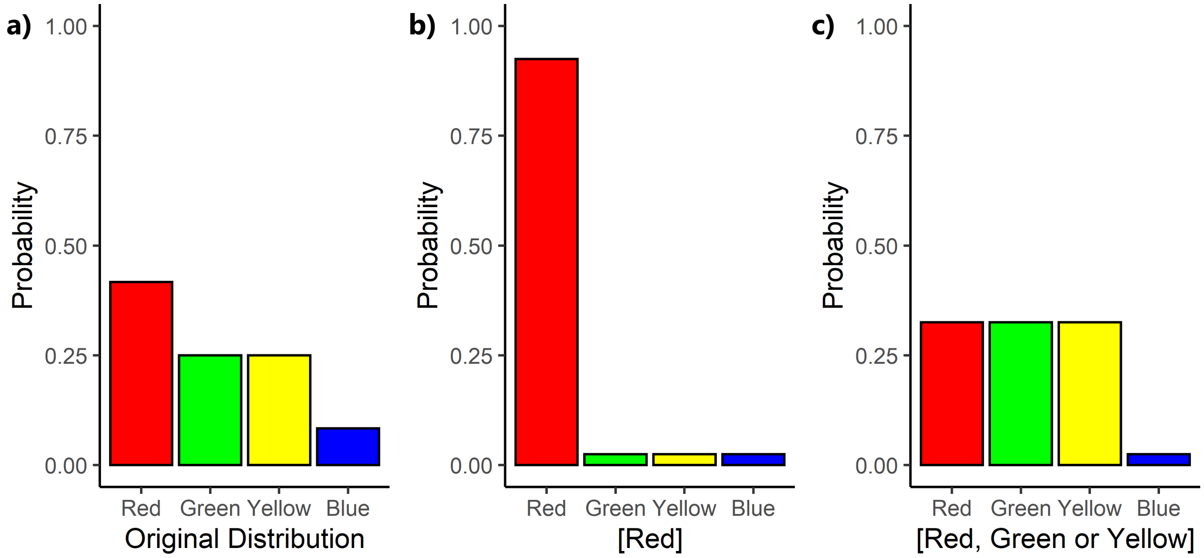

Figure 2: **a:** Original distribution over possible outcomes, in the speaker's mind. **b:** a distribution over possible outcomes consistent with the guess "it will be a red ball". **c:** a distribution over possible outcomes consistent with the guess "it will be a red, yellow or green ball".

comes it mentions), and n the number of possible outcomes[2]. λ is a free parameter that controls how much probability mass we want to put on possible outcomes that are not mentioned in the guess[3].

When is a guess good? Intuitively, it is when the probability distribution that the audience can construct from the guess is 'close' to the original distribution in the speaker's mind. For example, on Figure 2, the distribution in 2c looks 'closer' to the original distribution (2a) than the distribution in 2b does.

We can formalize this intuitive notion of 'closeness' between distributions using tools from information theory. Specifically, we say that a guess G is a good encoding of the original distribution Q if the Kullback-Leibler divergence (KL-D) of G from Q is low (see Gagie, 2006). Thus, we define the quality of a guess as the inverse of the KL-D of G from Q (we add 1 to the denominator so that guess quality can range from 0 to 1):

$$\frac{1}{1+KL(Q|G)} = \frac{1}{1+\Sigma_i Q(i)\log\left(\frac{Q(i)}{G(i)}\right)}$$

Finally, different people might prefer to make slightly different guesses, or they might make different guesses in different contexts. For instance, if you are in a hurry you might prefer to make a short guess, like "It will be red", but if you are very cautious you might say "It can be any of the four colors". It is not clear how to best model this variability, but here we assume that before making a guess, speakers modulate their original probability distribution by either concentrating it in a few peaks, or spreading it out more evenly. That is, the speaker constructs a modified distribution Q' by applying the following transformation to each element i of Q:

$$Q'(i) = \frac{Q(i)^\alpha}{Z}$$

where Z is a normalizing constant ensuring that all elements in Q' sum to 1, and α is a free parameter which controls to what extent the distribution gets concentrated or spread out. For values of $\alpha < 1$, the probability distribution gets spread out; for $\alpha > 1$, it gets concentrated (areas with a lot of probability mass get even more probability mass to the detriment of other areas). Low values of α result in guesses that mention more possible outcomes.

The accuracy-informativeness trade-off model

We also consider an alternative formal model of the psychology of guesses. This model (Dorst & Mandelkern, 2021) is based on the hypothesis that people make guesses that optimize a trade-off between accuracy and informativeness (see Yaniv & Foster, 1995, 1997). A guess is accurate if it is likely to be true, and it is informative if it gives information about which outcomes are particularly likely. For instance, the guess "the ball will be either Red, Green, Yellow or Blue" has probability 1, so it is very accurate. But it gives no information about whether some colors are more likely to come out. So, in practice, people might prefer other guesses, that

[2] In situations where we don't know the space of possible outcomes, we can for example assign probability $\frac{\lambda}{n_g+1}$ to the general probability that something not mentioned in the guess occurs.

[3] We find that the value of λ has little influence on the model's fit in the studies we report below. To limit the number of free parameters, we simply set $\lambda = .1$

are not guaranteed to be correct but that narrow down the space of plausible outcomes.

According to Dorst & Mandelkern (2021), people select the guess G that maximizes the following expected utility function:

$$E(G) = P(G) * J^{Q(G)}$$

where $P(G)$ is the *accuracy* of a guess, i.e. the probability that the guess is correct, and $Q(G)$ is its *informativeness*. The informativeness of a guess is the proportion of possible answers that it excludes. For example, the guess "red or blue" has informativeness $1/2$ because it excludes two of the four possible options. The parameter J regulates how sensitive people are to informativeness relative to accuracy (when $J = 1$, people are only sensitive to accuracy; when $J >> 1$ they mostly care about informativeness).

Note that our information-theoretic model is consistent with the hypothesis that people make guesses that achieve a trade-off between accuracy and informativeness. The difference is that our model does not assume that the mind is explicitly optimizing this trade-off – instead, the tradeoff happens as a natural byproduct of making guesses that efficiently encode the probability distribution in the speaker's mind. We also note that the two models make very similar predictions, and our main goal here was not to test which model fits people's intuitions better.

It is easy to see that our model, as well as the trade-off model, account for qualitative features of our intuitions about which guesses are good. For instance, our model explains why we think that 'it will not be red' is a worse guess than 'it will be red'. 'It will not be red' assigns a low probability to Red, which is actually the highest-probability outcome in the speaker's mind, and therefore it defines a distribution that does not look anything like the one in the speaker's mind.

In the following section, we test whether our account (and the trade-off model) can account for *quantitative* patterns in people's judgments about what counts as a good guess.

Methods

In two studies, we showed participants urns that were similar to the one shown in Figure 1, but whose content we systematically varied, in a within-subject design.

One natural way to test our hypothesis would be to ask people to make guesses about which ball will come out from a given urn. In order to obtain many data points per participants, we instead asked them to rate the quality of different guesses that one could make. We compared their ratings with the predictions of our information-theoretic model (henceforth, KL model), and the predictions of the accuracy / informativeness trade-off model.

Data and R code (for modeling and data analysis) are available on the OSF at `https://osf.io/wfgya/`.

Materials and Measures

Each participant saw 13 (in study 1) or 10 (in study 2) different urns, each containing 12 balls of different colors (Red, Yellow, Blue, Green; there was at least one ball of each color in each urn). To construct the urns, we defined the 'profile' of an urn as a list of four numbers, specifying the number of balls of the most frequent color, the number of balls of the second most frequent color, and so on. For example, the urn in Figure 1 has profile [5,3,3,1]. The content of the urns was procedurally generated. For each urn and each participant, the frequency ordering over colors was randomized. The generation process also randomized the position of the balls in a given urn, the order of presentation of urns, and the order of presentation of guesses. All guesses for a given urn were presented alongside the urn on a single page. Different urns were presented on different pages.

For each urn, we asked participants to rate the quality of four guesses, on a Likert scale from 1 (bad guess) to 9 (good guess). The guesses were of the form "The player will draw {}", where {} was a disjunction of possible colors (e.g. "a red ball or a yellow ball"). We call the number of colors in {} the size of a guess. For example, "Red or Yellow" is a guess of size 2. We constructed four guesses, of sizes 1, 2, 3 and 4, per urn, by first building a guess with the most frequent color, then a guess with the two most frequent colors, etc. For example, for the urn shown in figure 1, we constructed the guesses {Red}, {Red or Yellow}, {Red, Yellow or Green} and {Red, Yellow, Green or Blue} (In cases where some colors have equal frequency we randomly imposed an artificial ordering on them when constructing guesses).

Procedure

Participants were recruited on Prolific and completed the experiment on a web-based interface. We first asked participants to familiarize themselves with the setting by randomly drawing a few times from two different urns. Then they read a short set of instructions explaining the task. In the main phase of the study, participants rated the quality of four guesses per urn – each page featured a picture of a different urn, alongside four different guesses to rate. Participants then completed a short set of questions probing whether they understand how probability works in the current context (we do not analyze these reports here). Finally, they completed a few demographic questions and were redirected to Prolific for payment.

Studies 1 and 2 had essentially identical designs, with the following exceptions. Study 2 was shorter, with 10 instead of 13 different urns per participant. It also featured slightly different instructions. While in study 1 we simply told people that they were about to rate different possible guesses, in study 2 we asked them to imagine that they would be communicating with a friend who cannot see the contents of the box (but knows that boxes contain red, blue, green and yellow balls, in unknown proportion). Likert scales were labelled with 'bad guess' and 'good guess' in study 1, and 'bad answer' and 'good answer' in study 2.

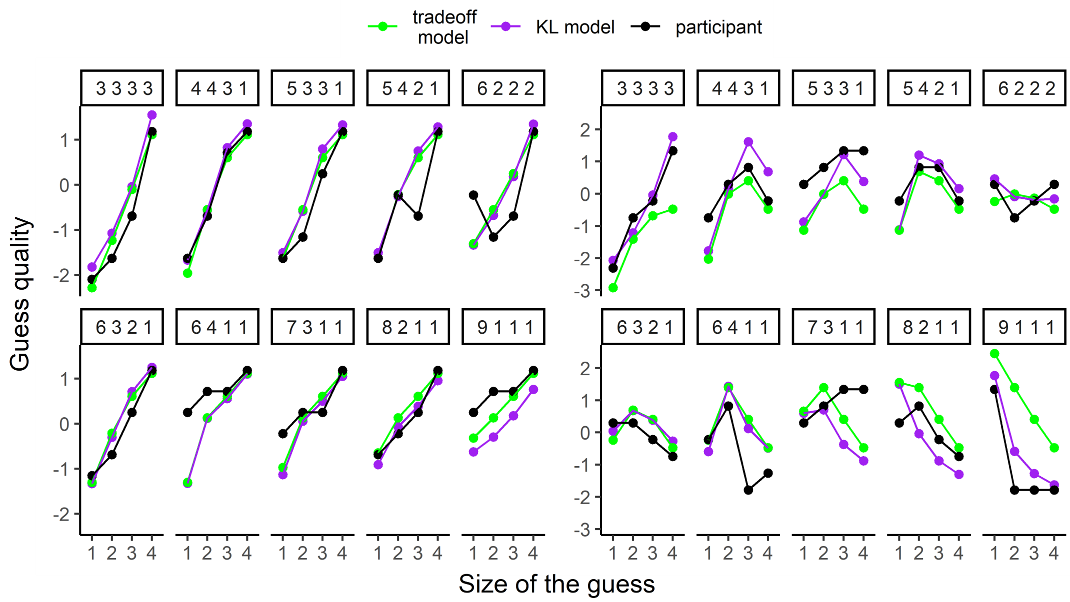

Figure 3: Ratings from two representative participants in Study 2, along with the predictions of the KL (purple) and trade-off (green) models. The participant on the left has a low best-fitting value of α, while the participant on the right has a high value of α. Panel labels represent the profile of an urn: for example, an urn labelled [9,1,1,1] has 9 balls of one color, and one ball each of the other colors. Participant and model ratings are z-scored to facilitate comparison.

Interested readers can walk through the experiments at https://eco.ppls.ed.ac.uk/~tquillie/guesses/ and https://eco.ppls.ed.ac.uk/~tquillie/guesses-b/.

Participants

We recruited US residents (in study 1, N=38, 24 female, 13 male, 1 other, mean age = 30.8, SD = 9.5; in study 2; N=39, 24 female, 14 male, 1 other, mean age = 30.7, SD = 9.4) from Prolific. Participants were compensated £1 for their participation.

Modeling

For each participant, we generate model predictions for the KL model by finding the value of α that maximized the correlation between model predictions and the participant's ratings, using the optim function in R (R Core Team, 2022). We generated model predictions for the tradeoff model in a similar way (fitting the J parameter individually to each participant).

Results

We find very similar results in both studies. Most participants adopted a simple strategy: they simply rated the quality of a guess according to its probability. That is, they gave highest ratings to guesses of size 4, i.e. guesses that mention all possible outcomes and therefore have probability 1. Figure 3 (left panel) shows the ratings made by one such participant (in study 2).

Both models are able to closely reproduce this pattern of judgments. The trade-off model can trivially explain the pattern: when setting J to 1, the model holds that people are only sensitive to accuracy – that is, the quality of a guess is just its probability. The KL model captures the pattern of judgments by setting $\alpha << 1$, i.e. by assuming that people compare a guess to a 'spread out' probability distribution which gives very similar probability to each possible outcome.

There was nonetheless also a substantial number of participants (in both studies) who exhibited a more subtle pattern of judgments – see for example the participant highlighted on the right of Figure 3. These participants tend to favor long guesses when colors are equally frequent (as in the urn with profile [3,3,3,3] which has 3 balls of each color), but they preferred shorter guesses for urns where one color was predominant. For example, for an urn with 9 red balls, these participants would favor the guess "The player will draw a red ball". For an urn with 6 yellow balls and 4 blue balls, many of them would favor the guess "The player will draw a yellow ball or a blue ball".

Again, both computational models are able to account for this pattern of judgments.

For values of α that are not too low, the KL model favors guesses which put a lot of probability mass on the most likely outcomes, because such guesses define distributions that are close to the actual probability distribution over possible outcomes. Therefore the model naturally favors short guesses when one or a few colors dominate, and long guesses when all colors have the same frequency.

The trade-off model accounts for the data because, for $J > 1$, it values guesses that are both likely and informative. For

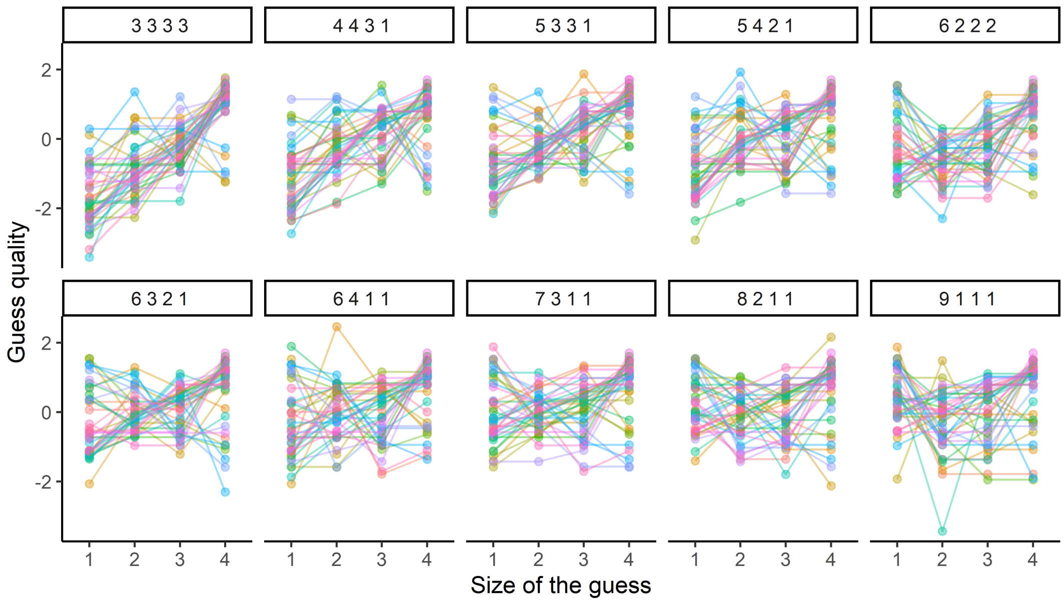

Figure 4: Ratings from all participants in Study 2. To make the trends easier to discern, the ratings of each participant are z-scored. Panel labels represent the profile of an urn: for example, an urn labelled [9,1,1,1] has 9 balls of one color, and one ball each of the other colors. A similar figure for Study 1 can be found at `https://osf.io/6utsk/`.

an urn with 9 red balls of the same color, the guess "it will be red" is likely enough (it will come out true 75% of the time), and it is very informative because it rules out 3/4 of the possible outcomes. By contrast, for an urn where all colors are present in equal proportion, leaving out one color from the guess decreases its probability and makes it more informative – but the gain in informativity is not large enough to be worth the bargain. Therefore the model favors long guesses, like "it can be any color".

We display the fit between individual participant judgments and model judgments, for both models and both studies, in Figure 5. We also show the fit of a simple model for which the quality of a guess is always its probability. The judgments of all participants are positively correlated with the predictions of both the KL and trade-off models.[4] The simple probability model, by contrast, provides a very bad account of the judgments of some participants.

Discussion

Probabilistic theories of cognition (Tenenbaum et al., 2011; Oaksford & Chater, 2007) assume that the mind often represents probability distributions over possible outcomes, or over possible states of the world. At first sight this assumption is at odds with the way people express uncertainty when they talk. If someone asks you how many people you think live in the European Union, you will probably not say "my subjective belief is a normal distribution with mean 300 million inhabitants and standard deviation 50 million" – instead you will probably say something like "between 200 and 400 millions".

Here we have sketched a theory that resolves this tension.[5] We assume that when people respond to a question whose answer they are uncertain of, their mind does represent a probability distribution over possible answers. But it would not be convenient for them to enumerate this distribution out loud. Instead, they make a simple guess, which effectively functions as a compressed representation of the probability distribution in their mind. Someone who hears the guess cannot reconstruct the exact probability distribution that the speaker had in mind, but can infer a good enough reconstruction.

Our account assumes that speakers choose messages that optimize the inferences they anticipate the listener will draw, in the spirit of recent computational models of social cognition (Goodman & Frank, 2016; Shafto et al., 2014; Franke & Jäger, 2016). In particular, we assume that the speaker wants the listener to re-construct a probability distribution that is close to the speaker's subjective distribution, as in recent models of how people use words such as "about" (Egré et al., 2020) and "possibly", "likely", or "almost certainly"

[4]Participants with a low best-fitting value of α are well-fit by both models, but the trade-off model provides a slightly better account of their judgments. This is not surprising given that the probability of a guess is a basic building block of the trade-off model: by setting $J = 1$, the trade-off model reduces to a model that judges that more probable guesses are always better. For participants with higher values of α, there is no clear advantage for one model over the other.

[5]There are other domains where it does not intuitively feel like the mind represents probability distributions. Perception is one example: we do not simultaneously 'see' all the possible interpretations of an ambiguous visual stimulus (like a Necker cube); see Gershman et al. (2012) for one possible explanation.

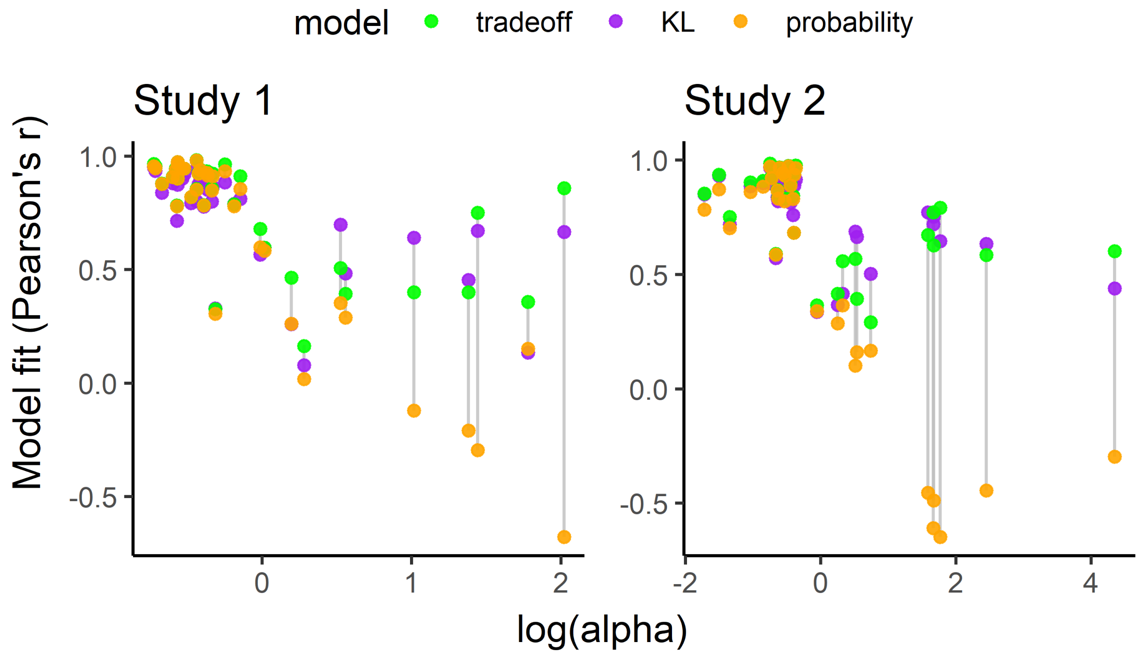

Figure 5: Individual model fits, studies 1 and 2. Each point corresponds to the correlation between the judgments of one participant and the trade-off model (green), the KL model (purple) or the simple probability model (orange). Gray lines connect points belonging to the same participant. The x-axis represents the best-fitting value of α for a given participant.

(Herbstritt & Franke, 2019) to communicate uncertainty.

In two studies, we show that the theory can account for people's judgments about what counts as a good guess in response to a question about an uncertain outcome.

Accuracy and informativeness

Our theory is consistent with the hypothesis that people make guesses that are both informative and likely to be correct (Yaniv & Foster, 1995, 1997; Dorst & Mandelkern, 2021). Indeed, our model generally favors guesses that have high probability, but that also convey some information about which outcomes are more likely than others. A trade-off between accuracy and informativeness thus emerges as a byproduct of making guesses that efficiently communicate the subjective probability distribution in the speaker's mind. In other words, the model explains, from first principles, why good guesses appear to be sensitive to the accuracy / informativeness trade-off.

Of course, an alternative explanation is that the mind actually computes explicit representations of the informativeness and accuracy of a guess. Dorst & Mandelkern (2021) have developed a model based on this assumption, and we find that it accounts to the present data as well as our theory does. Both models make quite similar predictions in the task we used here – critical tests between the two accounts are a ripe direction for future research.

Our information-theoretic approach can easily be applied to cases where the relevant probability distribution is continuous (for example, your subjective belief about the number of people living in the EU). The trade-off model could in principle also model such cases, although it might need a slightly different definition of informativeness in order to do so. Dorst & Mandelkern (2021) currently define the informativeness of a guess as the proportion of possible outcomes it does not mention, which can be ambiguous when the space of possible outcomes is continuous.

Conclusion

People can form complex representations of the world, and they face the challenge of efficiently communicating these representations to others (Grice, 1975; Sperber & Wilson, 1986; Pinker, 2007; Goodman & Frank, 2016). Cognitive scientists have found that the mind seems well-equipped to meet this challenge: people can often convey a surprising amount of information with very few words. For example, people can transmit a lot of statistical information about the properties of a category by using *generics*, such as 'robins lay eggs' (Tessler & Goodman, 2019). They are also able to communicate the content of their complex causal models of the world, by making short statements about what could have been (Lucas & Kemp, 2015), or highlighting one cause among the many factors that contributed to an event (Kirfel et al., 2021; Quillien, 2020). Here, we suggested that people can also also efficiently express their probabilistic beliefs about the world in the form of simple guesses.

Acknowledgments

We thank three anonymous reviewers, and members of the Edinburgh Computational Cognitive Science Laboratory, for helpful feedback. This research was supported by an EPSRC New Investigator Grant (EP/T033967/1) to C.L.

References

Dorst, K., & Mandelkern, M. (2021). Good guesses. *Philosophy and Phenomenological Research*. doi: 10.1111/phpr.12831

Egré, P., Spector, B., Mortier, A., & Verheyen, S. (2020). On the optimality of vagueness: "around", "between", and the gricean maxims. *arXiv preprint arXiv:2008.11841*.

Franke, M., & Jäger, G. (2016). Probabilistic pragmatics, or why bayes' rule is probably important for pragmatics. *Zeitschrift für sprachwissenschaft*, *35*(1), 3–44.

Gagie, T. (2006). Compressing probability distributions. *Information Processing Letters*, 133–137.

Gershman, S. J., Vul, E., & Tenenbaum, J. B. (2012). Multistability and perceptual inference. *Neural computation*, *24*(1), 1–24.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, *20*(11), 818–829.

Grice, H. P. (1975). Logic and conversation. In *Speech acts* (pp. 41–58). Brill.

Herbstritt, M., & Franke, M. (2019). Complex probability expressions & higher-order uncertainty: Compositional semantics, probabilistic pragmatics & experimental data. *Cognition*, *186*, 50–71.

Holguin, B. (2022). Thinking, guessing and believing. *Philosopher's Imprint*, 1–34.

Kirfel, L., Icard, T., & Gerstenberg, T. (2021). Inference from explanation. *Journal of Experimental Psychology: General*.

Lucas, C. G., & Kemp, C. (2015). An improved probabilistic account of counterfactual reasoning. *Psychological review*, *122*(4), 700.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.

Pinker, S. (2007). *The stuff of thought: Language as a window into human nature*. Penguin.

Quillien, T. (2020). When do we think that x caused y? *Cognition*, *205*, 104410.

R Core Team. (2022). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive psychology*, *71*, 55–89.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition* (Vol. 142). Citeseer.

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.

Tessler, M. H., & Goodman, N. D. (2019). The language of generalization. *Psychological review*, *126*(3), 395.

Yaniv, I., & Foster, D. P. (1995). Graininess of judgment under uncertainty: An accuracy-informativeness trade-off. *Journal of Experimental Psychology: General*, *124*(4), 424.

Yaniv, I., & Foster, D. P. (1997). Precision and accuracy of judgmental estimation. *Journal of behavioral decision making*, *10*(1), 21–32.