Using alien coins to test whether simple inference is Bayesian

Peter Cassey

University of Newcastle

Guy E. Hawkins

University of Amsterdam

Chris Donkin

University of New South Wales

Scott D. Brown

University of Newcastle

Author Note

Peter Cassey, School of Psychology, University of Newcastle; Guy E. Hawkins, Amsterdam Brain and Cognition Center, University of Amsterdam; Chris Donkin, School of Psychology, University of New South Wales; Scott D. Brown, School of Psychology, University of Newcastle

This research was supported in part by Australian Research Council Future Fellowship FT120100144. P. Cassey and S. D. Brown developed the study concept. All authors contributed to the study design. Testing and data collection were performed by P. Cassey. All authors performed the data analysis and interpretation. All authors drafted the manuscript. All authors approved the final version of the manuscript for submission.

Correspondence concerning this article should be addressed to Peter Cassey, School of Psychology, University of Newcastle, University Drive, Callaghan, NSW, Australia, 2308. peter.cassey@newcastle.edu.au

Abstract

Reasoning and inference are well-studied aspects of basic cognition that have been explained as statistically optimal Bayesian inference. Using a simplified experimental design, we conducted quantitative comparisons between Bayesian inference and human inference, at the level of individuals. In three experiments, with more than 13,000 participants, we asked people for prior and posterior inferences about the probability that one of two coins generated certain outcomes. Most participants' inferences were inconsistent with Bayes rule. Only in the simplest version of the task did the majority of participants adhere to Bayes rule, but even in that case there was a significant proportion that failed to do so. The current results highlight the importance of close quantitative comparisons between Bayesian inference and human data at the individual-subject level when evaluating models of cognition.

Keywords: Bayesian inference; inference; prior knowledge

Using alien coins to test whether simple inference is Bayesian

Bayesian inference is a mainstay of modern statistical analysis, but it has also become influential as a description of human cognition. Bayesian belief updating involves two elements: a prior, which represents belief states before observing data, and a likelihood function, which links observed evidence with beliefs by assigning probabilities. The likelihood and the prior are combined (via Bayes rule) to give an updated belief, the posterior. Among other aspects of human cognition, Bayesian models have provided compelling explanations for language acquisition (Griffiths & Kalish, 2007), language evolution (Maurits & Griffiths, 2014; Rafferty, Griffiths, & Klein, 2014), word learning (Xu & Tenenbaum, 2007), speech recognition (Norris & McQueen, 2008), reading (Norris, 2006), causal learning (Griffiths & Tenenbaum, 2009), cultural transmission (Kalish, Griffiths, & Lewandowsky, 2007), future prediction (Griffiths & Tenenbaum, 2006, 2011; Lewandowsky, Griffiths, & Kalish, 2009), and visual working memory (Brady & Tenenbaum, 2013).

Despite – or perhaps, because of – their success, Bayesian models have sparked some criticism. Specific models have been criticized for complexity, which may reduce their explanatory power. Mozer, Pashler, and Homaei (2008) demonstrated how a model based on Bayesian inference was not necessary to account for participants' behavior on a future prediction task (Griffiths & Tenenbaum, 2006). Mozer et al.'s simplified heuristic model performed with commensurate success to that of the Bayesian model, questioning the necessity of the substantial addition of theory (however, Lewandowsky, Griffiths, & Kalish, 2009, identified important problems with the simplified model).

Similar results in other paradigms have led to more general debates about the role of Bayesian models in human cognition (cf. Bowers & Davis, 2012a; Bowers & Davis, 2012b;

Chater et al., 2011; Eberhardt & Danks, 2011; Griffiths, Chater, Norris, & Pouget, 2012; Griffiths, Vul, & Sanborn, 2012; Jones & Love, 2011; Marcus & Davis, 2013). These debates have included the value of normative models, with philosophical arguments about the role of explanations posed at Marr's algorithmic versus computational levels. More tangibly, the choice of priors and likelihood functions have been criticized as conferring undue model flexibility. In a Bayesian model of cognition, changes in the prior lead to changes in the model's predictions. This can be problematic because there are situations in which the prior that participants really use is difficult or impossible to ascertain (however, see Hemmer, Tauber, & Steyvers, 2014). The degree to which Bayesian models of cognition are quantitatively tested against human data has also been highlighted as a limiting factor (Hemmer et al.).

Our research was designed to overcome this problem, and others, by designing the experimental paradigm to allow easy communication and measurement of prior and posterior beliefs. We quantitatively examined the degree to which human behavior approximated Bayesian inference at the level of the individual subject, using the well-studied paradigm of simple probabilistic inference. While previous research has demonstrated some agreement between posterior probability distributions from Bayesian inference and from people, these tests have most frequently been applied at the group level (Griffiths & Tenenbaum, 2006, 2009, 2011; Lewandowsky et al., 2009; Shi, Griffiths, Feldman, & Sanborn, 2010). When these investigations have been applied at the level of individuals, the comparison between human and Bayesian inference has been largely qualitative (e.g. Williams & Griffiths, 2013); that is, analysis questions are frequently of the type "do the participants' responses move in the direction predicted by Bayesian inference". Our experimental paradigm reduced the inference task to a level that allowed the Bayesian model of cognition to be quantitatively compared with individual

participants' behavior. We manipulated the difficulty of a simple prediction about an alien who was flipping coins. Participants were asked about the nature of a coin (i.e., fair or biased) both before and after seeing a sequence of outcomes. We recruited a sufficiently large number of participants that the full ranges of prior and posterior beliefs were sampled, and also enough that we were able to analyze important subsets of the data. The participants' inferences were mostly inconsistent with Bayes rule. However, as the prediction scenario became simpler, more participants responded in a manner that was consistent with Bayesian inference.

Experiment 1

Method

Participants. For each experiment reported we collected data from 4,000-5,000 participants because this provided sufficient resolution to calculate the density of responses in a 17x17 grid of prior vs. posterior probabilities. For Experiment 1, 4,033 USA-based participants were recruited online via Amazon's Mechanical Turk. The experiment took an average of three minutes to complete and participants were paid USD$0.50.

Procedure. At the start of the experiment participants were told to imagine that they were on an alien planet called "Cointopia", where only two types of coins exist. One type of coin was unbiased, like our earth coins (called a "zonk"). The other type of coin was biased such that there was a 70% chance of head and a 30% chance of a tail (called a "zlink").

Participants were next told that they met a local of Cointopia, an alien called "Zed", who was holding a coin, but they did not know what type of coin. Participants were instructed to move a slider to indicate the probability (in percent) that Zed was holding a zonk. The slider was bounded at 0 and 100. A description of what a response of 0, 50 and 100 meant was provided

above the slider (e.g., "A slider all the way to the left (0) indicates that you believe there is 0% probability that Zed has a zonk. This means that that you believe there is 100% probability that Zed has a zlink."). For convention, all analyses reported use the probability scale of [0,1] rather than percentages.

Participants then saw the result of four coin flips. Participants were randomly allocated to one of five outcome conditions; 0-, 1- , 2- , 3- or 4-tails. In each condition, participants randomly saw one of all possible orders of outcomes. For example, participants in the 1-tail condition saw either *T, H, H, H* or *H, T, H, H* or *H, H, T, H* or *H, H, H, T*. After this, participants were instructed to move the same slider to indicate the probability that Zed was holding a zonk. The same descriptions of what slider points 0, 50, and 100 meant were provided. This was where participants provided their posterior probabilities.

Results

Any participant who took less than one minute or more than 15 minutes to complete the experiment was removed from analysis due to considerations of engagement. These criteria removed 10% of participants.

The top row of Figure 1 plots participants' posterior probability estimates against their prior probability estimates. For this figure, both prior and posterior probability estimates have been binned into 17 ranges, and the number of participants falling into each bin is indicated by the size of the square in the plot. Just over half of the participants (55%) provided a prior probability of exactly 0.5, with 61% of participants giving a prior probability between 0.45 and 0.55.

The solid black lines in Figure 1 indicate, for any given prior, the Bayesian posterior. If participants had updated their prior in light of the coin flip outcomes exactly according to Bayes rule, then all data would lie on the black lines. That is, for any prior (x-axis value) the only square would be centered on the black line, and all other regions would be empty. To allow for some noise, we defined a posterior probability estimate as "Bayes optimal" if it was within 10 percentage points of the actual Bayesian posterior. These regions are illustrated by dashed lines in Figure 1. In Experiment 1 (top row), it is clear that many participants gave posterior probability estimates that were inconsistent with Bayes rule. Indeed, across Experiment 1, only 33% of participants provided posterior estimates that were within $\pm 10\%$ of the Bayesian value. This is a very low proportion, given that random uniform responses would lead to 20% (or just under, due to edge effects).

Possible explanations for the sub-optimal inferences we observed are that participants were confused by the task, or inadequately engaged in the task. To investigate these, we examined a subset of participants who seemed least likely to be disengaged; those who gave prior probability estimates close to 50% (we defined "close" as within the interval [0.45,0.55]). These participants actively moved the prior probability slider away from its random starting point, to indicate no strong prior beliefs about coins on an alien planet. We also addressed the possibility that some participants might have mixed up the polarity of the slider, despite the reminders, by re-assigning the posterior estimate of any participant who moved their posterior in the opposite direction from their prior estimate compared with the Bayesian posterior. These posterior estimates were reassigned using $\mathrm{p} \mapsto (1 - \mathrm{p})$. We return to the issue of participant engagement in the discussion.

The resulting distributions of posterior probability estimates are shown in Figure 2. The gray regions capture responses that were within $\pm 10\%$ of the Bayesian posterior corresponding to a prior of 0.5. The percentage of participants within 10% of the Bayesian posterior is displayed above each panel. Around 47% of participants gave posterior probability estimates within 10% of the Bayesian posterior. However, even uniform random responding would lead to 30% of participants falling within 10% of the Bayesian posterior, by chance (more than 20%, due to the generous re-assignment $p \mapsto (1 - p)$ for any participant who updated their prior in the wrong direction). Thus, the performance of the participants in Experiment 1 was certainly different from Bayesian inference, even when we made considerations for confusion about the scale polarity, or lack of engagement with the task.

Discussion

Overall, inferences in Experiment 1 were inconsistent with Bayes rule. While some participants shifted their beliefs in the wrong direction, the inconsistency with Bayes rule was still evident even when all responses from these participants were given the benefit of the doubt, and interpreted as response polarity confusions. The results also cannot be explained by the well-known phenomenon of conservatism in belief updating; that is, the idea that people shift their beliefs more slowly, or by a lesser amount, than is optimal. In Experiment 1, two of the five outcome conditions led to the opposite of conservatism; there was *over*adjustment of beliefs when participants saw either one out of four tails, or two out of four tails (second and third panels from the left in the top row of Figures 1 and 2).

Experiment 2

Experiment 2 replicated Experiment 1 but with one new element: before providing posterior probability estimates, participants were reminded of the prior probability estimate they gave. We reasoned that this might help to reduce confusion and memory load.

Method

Participants. 5,015 USA-based participants were recruited online via Amazon's Mechanical Turk. The experiment took on average three minutes to complete and participants were paid USD$0.50.

Procedure. Identical to Experiment 1 in all aspects except that participants were reminded of their prior when providing their posterior, with the following text:

The value the slider is already on indicates your belief about the probability that Zed had a zonk before you saw the results of the coin flips.

Results

Following the same time latency exclusions as Experiment 1, 4% of participants were removed from analyses. The results of Experiment 2 are shown in the middle row of Figure 1. The results restricted to those participants who provided a prior probability in the interval [0.45,0.55], with allowance given for polarity confusion, are shown in the middle row of Figure 2. In both analyses, the results of Experiment 2 were very similar to the results from Experiment 1. Only 47% of participants provided a posterior probability estimate that was within $\pm 10\%$ of the Bayesian posterior, even when considering only those participants who provided a prior near 50%, and even allowing for confusion about slider polarity for any participant who moved their

posterior probability in the opposite direction from their prior compared to the Bayesian optimal posterior probability.

Discussion

The results from the first two experiments were extremely similar, shown in the top and middle rows of Figures 1 and 2. The two experiments even show the same pattern of conservative belief updating for participants who saw four tails from four coin flips, and the opposite effect for participants who saw one or two tails from four coin flips. These patterns suggest that the over- and under-updating of beliefs is a robust pattern in this paradigm, and not explained by simple effects such as anchoring due to the initial position of the slider used to indicate posterior probability. This slider was positioned at the prior probability estimate in Experiment 2, but not in Experiment 1, which might have been expected to induce greater conservatism in Experiment 2.

Experiment 3

Given the suboptimal inferences made by participants in Experiments 1 and 2, we wondered if optimal inference might be elicited if the inference problem was made very easy. In the first two experiments, the hypotheses (coins) were asymmetric: one was 50/50, the other 70/30. This might present a more difficult inference problem, because most observable evidence patterns are more likely under the unbiased coin than the biased coin. Experiment 3 made the inference problem easier, by using symmetric coins.

Method

Participants. 5,116 USA-based participants were recruited online via Amazon's Mechanical Turk. The experiment took on average three minutes to complete and participants were paid USD$0.50.

Procedure. Experiment 3 procedure was identical to Experiment 2 in all aspects except that the head/tail probabilities of the coins in Experiment 3 were symmetric. When participants received the initial scenario they were told that one type of coin (called a "zonk") was biased so that for any coin toss there was a 30% chance of getting a head and a 70% chance of getting a tail. The other type of coin (called a "zlink") was biased so that for any coin toss there was a 70% chance of getting a head and a 30% chance of getting a tail.

Results

Following the same time latency exclusions as Experiments 1 and 2, 5% of participants were removed from analyses. The bottom row of Figure 1 displays raw data from Experiment 3. Participants' prior probability estimates were closer to 50% than in Experiments 1 and 2, with 78% of participants providing a prior probability in the interval [0.45, 0.55]. The bottom row of Figure 2 shows the posterior probability estimates from Experiment 3 with the same data filtering as before: restricted to participants who gave a prior probability between 45% and 55%, and also giving the benefit of the doubt to any participant who updated their posterior in the wrong direction, and thus may have been confused about the response slider's polarity. This time, around 63% of participants provided posterior probability estimates within $\pm 10\%$ of the Bayesian optimal posterior (the chance level for this analysis is 25%, due to boundary restrictions on the 4-tail and 0-tail conditions). When participants were shown symmetric outcomes (2 heads and 2 tails), 89% provided posterior probabilities within $\pm 10\%$ of the

Bayesian optimal value. This is perhaps unsurprising: if one's prior probability estimate is 50%, and the data observations contain 50% heads vs. tails, then the appropriate posterior probability is also 50%. Across the other four heads/tails conditions, performance was much poorer, and closer to that observed in Experiments 1 and 2, with fewer than 55% of participants providing close-to-Bayesian posterior probabilities.

Discussion

The reduction in apparent complexity of Experiment 3 resulted in more participants providing close-to-Bayes-optimal inferences, particularly in the very easiest condition. Across the conditions, however, nearly one participant in three deviated by more than 10% from Bayes-optimality, and this held even after generous data filtering was applied in favor of observing optimality. Further, a comparison of the top two rows of Figure 1 against the bottom row shows that participants were mostly insensitive to the important difference in the hypotheses being compared. In all conditions except for the 2-head-2-tail condition, participants drew very similar inferences in Experiment 3 as in Experiments 1 and 2, even though this was not justified (note the differences between the Bayesian predictions).

General Discussion

Bayesian inference has been proposed as an analogy for human cognition in some paradigms. However, there has been recent and growing debate about the framework in general. Rather than engage in such general debates, we opted for a specific quantitative test of the correspondence between human experimental data and the predictions that come from belief updating via Bayes rule. We used a simple inference task in which participants judged what type of coin was likely to have generated a given series of outcomes. A key advance of our task was that it supported

precise quantitative comparisons between the posterior probabilities provided by Bayesian inference and those provided by participants. With more than 13,000 participants across three experiments, inferences were mostly inconsistent with Bayes rule.

It is well documented that humans have a propensity to discount the value of initial information in favor of novel information. With regards to probabilities, it has been shown that people often underweight initial beliefs and overweight new information, a phenomenon termed base–rate neglect, or insensitivity to the prior (Bar-Hillel, 1980; Tversky & Kahneman, 1974). The opposite phenomenon has also been observed in probability judgments; namely, over-weighting prior beliefs, and updating them more slowly than demanded by data (for review, see Weber, 1994). On their own, neither of these phenomena can explain the results of any of our three experiments, because we observed both over- and under-weighting of the prior probability, across different conditions. The two opposing phenomena could, together, explain the results, but only in a rather unsatisfying, post hoc manner.

In contrast to our results, basic human inference has previously been framed as statistically optimal Bayesian inference. Williams and Griffiths (2013) used a task with many similarities to ours, but found evidence in favor of a Bayesian interpretation of cognition. They presented participants with a sequence of coin flip outcomes that were generated by one of two coins differing in probability (just like our experiments). Knowing what these probabilities were, participants were asked to indicate which of the two coins generated the sequence, and their responses agreed overwhelmingly with the optimal Bayesian response.

Our experiments are similar to Williams and Griffiths' (2013), but our results are apparently very different: most of our participants deviated substantially from Bayesian optimal responses. The key to explaining this difference is our use of a more fine-grained response

measure, and hence more fine-grained comparison between humans and Bayes, than did Williams and Griffiths. Our participants provided a quantitative indication of their degree of belief in one coin over the other as opposed to a qualitative preference between two coins, whereas Williams and Griffiths' participants provided only a qualitative indication of which coin was more likely. We confirmed that this difference in response measure was a likely cause of the difference in results by discretizing our data to match the qualitative nature of Williams and Griffiths' data. We inferred each participants' preference for the two coins by assuming that a posterior > .5 indicated a preference for a *zonk* and a posterior < .5 indicated a preference for a *zlink* (if they were to make a forced two alternative choice). These inferred choices overwhelmingly matched the Bayesian optimal choices, just like Williams and Griffiths found: the match rate was 90% for Experiments 1 and 2, and 96% for Experiment 3. This analysis highlights the importance of testing cognitive theories at a quantitative level. The Bayesian theory of cognition, which is apparently successful when tested at a qualitative level, fails when tested quantitatively.

A key advance of our research was the ability to make precise and quantitative comparisons of posterior probability estimates against Bayes-optimal posterior probabilities, at a single-participant level. This advance was made possible by restricting the inference problem given to participants to a very simple situation with just two possible hypotheses – the alien could be holding one of only two coin types that were available. This restriction allowed an individual person's prior and posterior probabilities to be conveyed by just one number: the probability that the alien was holding one of the two coin types. We hoped that this representation of the problem in our experiments' procedures agreed with participants' internal representation of the problem, but this may not have been the case. For example, an alternative

assumption is that participants represented the problem as a hierarchical statistical problem, in which case their prior might be better imagined as a distribution over all probabilities in the unit interval. The distinction between this assumption and our procedure can be made clearer by an example. Our procedure assumed that a participant might represent their prior knowledge with a statement like "there is a 53% probability that the alien holds a *zlink* coin". The hierarchical version might instead state: "there is an 8% probability that the probability that the alien holds a *zlink* coin is somewhere between 0% and 10%, and a 21% probability that the alien holds a *zlink* coin is somewhere between 10% and 20%, ...". It is not clear, to us at least, how to approach the problem of deciding which statistical framework the participants were using, and so we have chosen to base our analyses on the simplest assumption.

Our experiments all utilized data collection through the online labor market place of Amazon's Mechanical Turk. The viability of online data collection has been a subject of investigation for more than a decade already (e.g. Reips, 2001; Stanton, 1998; Topp & Pawloski, 2002). There is evidence to suggest that there is little difference between the quality of data collected on MTurk versus in-lab (Buhrmester, Kwang & Gosling, 2011; Gosling, Vazire, Srivastava, & John, 2004; Paolacci, Chandler, & Ipeirotis, 2010). In addition, numerous benchmark findings have been replicated using MTurk, across varied paradigms and research domains (Berinsky, Huber, & Lenz, 2012; Crump, McDonnell, & Gureckis, 2013; Paolacci et al., 2010; see Rand, 2012, for a review).

Notwithstanding the above reassurances, the issue of participant engagement is pertinent in the experiments reported here. One basic indicator of engagement is the experiment completion times of participants. Our experiment was very simple, and could be reasonably completed in a matter of minutes; in fact the average completion time was three minutes.

However, we considered completion times of less than a minute unreasonable as the participant would not have time to read all the necessary instructions. Likewise participants who took longer than 15 minutes were likely to have not been solely engaged in the task from start to finish. These criteria removed only 6% of participants, with no impact on our results or conclusions.

In addition to response latency considerations of engagement, the above replication of Williams and Griffiths' (2013) analyses is telling. That analysis indicated that over 90% of our participants provided responses that were qualitatively consistent with Bayesian optimality (even though they were quantitatively not). This agreement would be very unlikely if there was large-scale disengagement with the task.

As a final point, it is worth noting that the issue of engagement as it relates to optimal Bayesian decision-making is more complex than first glance may suggest. Our exclusion criteria targeted people who behaved randomly. As such, some of the participants in our experiment may have been using relatively little cognitive effort to update their beliefs in light of the data. However, theories of Bayesian decision-making posit that the optimality occurs automatically. In one sense, this must be true, since people are demonstrably poor at yielding optimal Bayesian solutions to problems they are asked to solve more analytically (e.g., Hawkins, Hayes, Donkin, Pasqualino, & Newell, in press). It may be that stricter exclusion criteria would yield a greater proportion of optimal responses, but then one need posit an inverted U-shaped relationship between cognitive effort and the production of optimal responses. We leave the testing of such a hypothesis to future work.

Conclusions

In three experiments, most people provided inferences that were inconsistent with a Bayesian account of cognition. In all but the easiest condition of the easiest experiment, fewer than two thirds of participants provided a close-to-Bayesian posterior probability, even after applying data filtering methods that were in favor of observing optimality. In the easiest condition of the easiest experiment, nearly 90% of participants provided close-to-Bayesian inferences (after filtering). More generally, these results open questions about the extent to which previous evidence in favor of Bayesian accounts of cognition were biased either by the use of very easy inference problems, or by the absence of close quantitative comparisons between Bayesian inference and human data at the individual-subject level.

References

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica, 44*, 211-233.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*, *20*, 351-368.

Bowers, J., & Davis, C. (2012a). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, *138*, 389-414.

Bowers, J., & Davis, C. (2012b). Is that what Bayesians believe? Reply to Griffiths, Chater, Norris, and Pouget (2012). *Psychological Bulletin*, *138*, 423-426.

Brady, T. F., & Tenenbaum, J. B. (2013). A probabilistic model of visual working memory: Incorporating higher order regularities into working memory capacity estimates. *Psychological Review*, *120*, 85-109.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*, 3-5.

Chater, N., Goodman, N., Griffiths, T. L., Kemp, C., Oaksford, M., & Tenenbaum, J. B. (2011). The imaginary fundamentalists: The unshocking truth about Bayesian cognitive science. *Behavioral and Brain Sciences* (commentary), *34*, 194-196.

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS ONE*, *8*, e57410.

Eberhardt, F., & Danks, D. (2011). Confirmation in the cognitive sciences: The problematic case of Bayesian models. *Minds and Machines*, *21*, 389-410.

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist*, *59*, 93.

Griffiths, T. L., Chater, N., Norris, D., & Pouget, A. (2012). How the Bayesians got their beliefs (and what those beliefs actually are): Comment on Bowers and Davis (2012). *Psychological Bulletin*, *138*, 415-422.

Griffiths, T. L., & Kalish, M. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science*, *31*, 441-480.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, *17*, 763-773.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, *116*, 661-716.

Griffiths, T. L., & Tenenbaum, J. B. (2011). Predicting the future as Bayesian inference: People combine prior knowledge with observations when estimating duration and extent. *Journal of Experimental Psychology: General*, *104*, 725-743.

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, *21*, 263-268.

Hawkins, G. E., Hayes, B. K., Donkin, C., Pasqualino, M., & Newell, B. R. (in press). A Bayesian latent-mixture model analysis shows that informative samples reduce base-rate neglect. *Decision*.

Hemmer, P., Tauber, S., & Steyvers, M. (2014). Moving beyond qualitative evaluations of Bayesian models of cognition. *Psychonomic Bulletin & Review, 22*, 1-15.

Jones, M., & Love, B. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, *34*, 169-231.

Kalish, M., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review*, *14*, 288-294.

Lewandowsky, S., Griffiths, T. L., & Kalish, M. (2009). The wisdom of individuals: Exploring people's knowledge about everyday events using iterated learning. *Cognitive Science*, *33*, 969-998.

Marcus, G. F., & Davis, E. (2013). How robust are probabilistic models of higher-level cognition? *Psychological Science*, *24*, 2351-2360.

Maurits, L., & Griffiths, T. L. (2014). Tracing the roots of syntax with Bayesian phylogenetics. *Proceedings of the National Academy of Sciences of the United States of America*, *111*, 13576-13581.

Mozer, M., Pashler, H., & Homaei, H. (2008). Optimal predictions in everyday cognition: The wisdom of individuals or crowds? *Cognitive Science*, *32*, 1133-1147.

Norris, D. (2006). The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review*, *113*, 327-357.

Norris, D., & McQueen, J. M. (2008). Shortlist b: A Bayesian model of continuous speech recognition. *Psychological Review*, *115*, 357-395.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*, 411-419.

Rafferty, A. N., Griffiths, T. L., & Klein, D. (2014). Analyzing the rate at which languages lose the influence of a common ancestor. *Cognitive Science*, *38*, 1406-1431.

Rand, D. G. (2012). The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology*, *299*, 172-179.

Reips, U. D. (2001). The Web experimental psychology lab: Five years of data collection on the Internet. *Behavior Research Methods, Instruments, & Computers*, *33*, 201-211.

Shi, L., Griffiths, T. L., Feldman, N., & Sanborn, A. (2010). Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin & Review*, *17*, 443-464.

Stanton, J. M. (1998). An empirical assessment of data collection using the Internet. *Personnel Psychology*, *51*, 709-725.

Topp, N. W., & Pawloski, B. (2002). Online data collection. *Journal of Science Education and Technology*, *11*, 173-178.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124-1131.

Weber, H. U. (1994). From subjective probabilities to decision weights: The effect of asymmetric loss functions on the evaluation of uncertain outcomes and events. *Psychological Bulletin*, *115*, 228-242.

Williams, J. J., & Griffiths, T. L. (2013). Why are people bad at detecting randomness? A statistical argument. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *39*, 1473-1490.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*, 245-272.

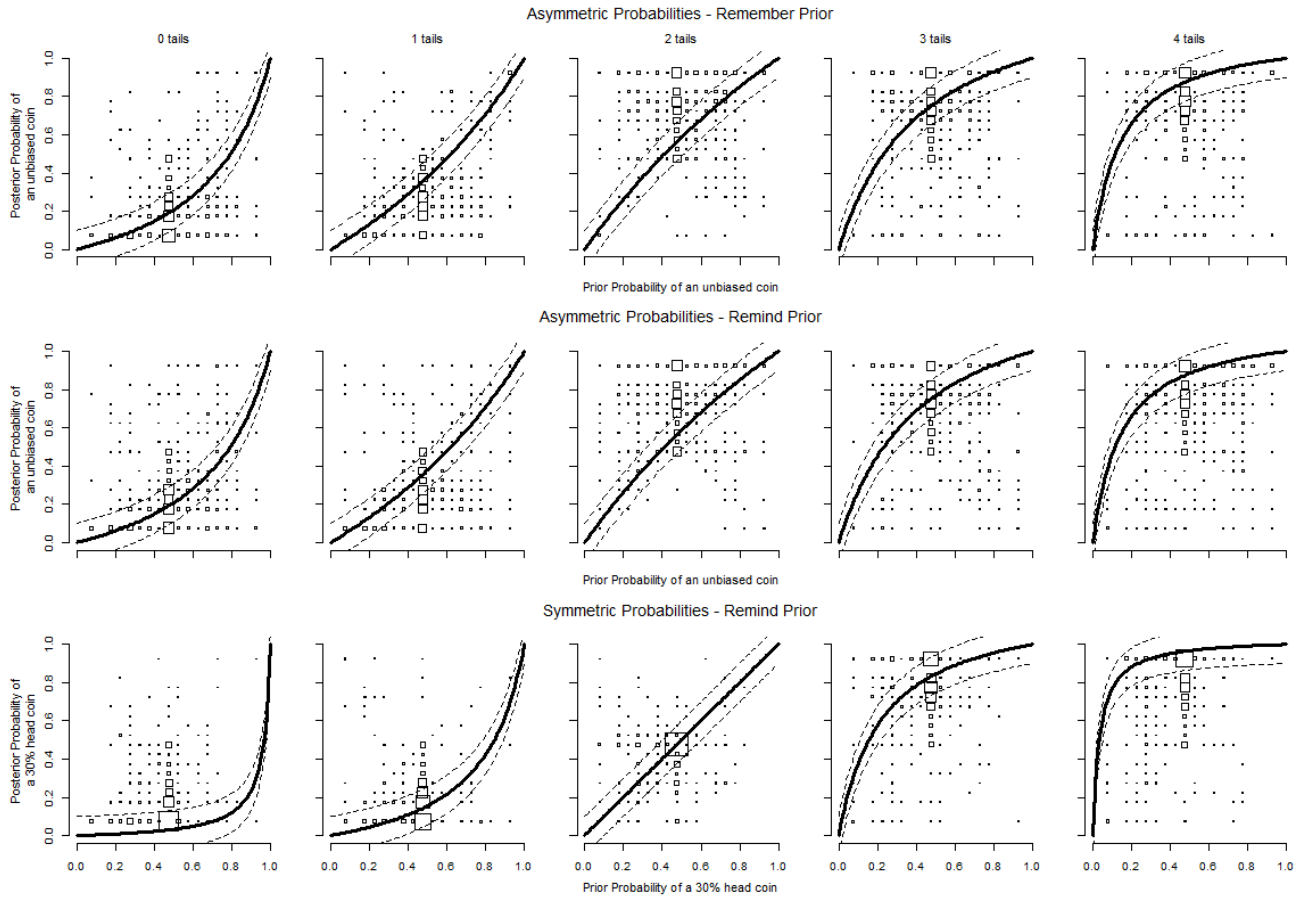

Figure 1. Posterior probability (y-axis) as a function of prior probability (x-axis). Larger squares indicate greater numbers of participants. Solid lines indicate the Bayesian optimal response conditioned on the prior probability. Dashed lines indicate posteriors that are within $\pm 10\%$ of the Bayesian posterior. The three rows correspond to three experiments, and the five columns correspond to the different coin flip outcomes (1 tail, 2, tails, etc.) that participants observed.

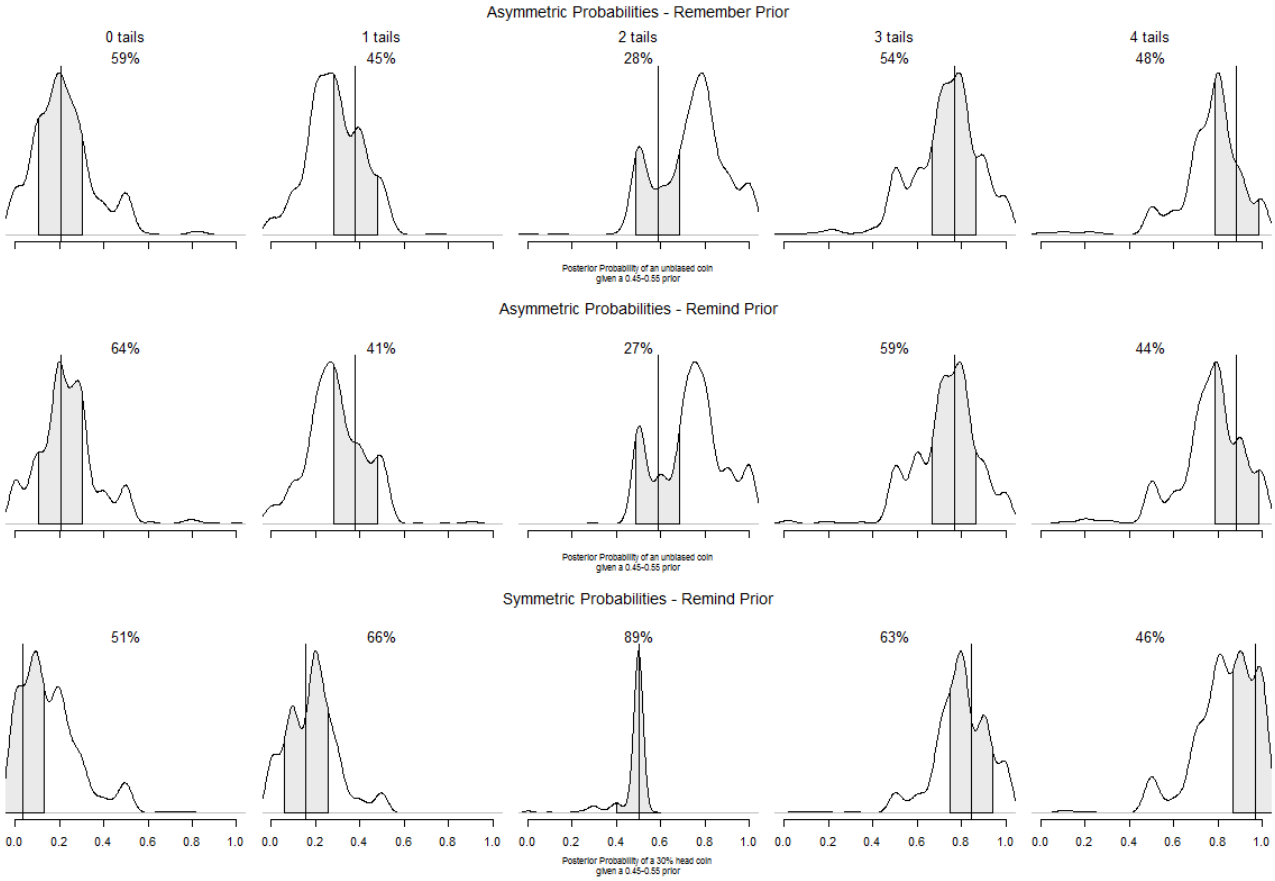

Figure 2. Distributions of posterior probability estimates for Experiments 1-3 (rows), restricted to those participants who provided a prior in the interval [0.45,0.55]. Any participant who moved their posterior in the opposite direction from their prior estimate compared with the Bayesian posterior was reassigned using $p \mapsto (1-p)$. Solid black line indicates the Bayesian optimal posterior given a prior of 50%. Shaded grey area indicates the optimal region, defined as 10% on either side of the true posterior. Values above each panel indicate the percentage of participants who reported a posterior within $\pm 10\%$ of the Bayesian value. Columns show N-Tail conditions.