Discovering simple heuristics from mental simulation

Frederick Callaway (fredcallaway@berkeley.edu)
Jessica B. Hamrick (jhamrick@berkeley.edu)
Thomas L. Griffiths (tom_griffiths@berkeley.edu)
Department of Psychology; University of California, Berkeley; Berkeley, CA, USA

Abstract

In the history of cognitive science, there have been two competing philosophies regarding how people reason about the world. In one, people rely on rich, generative models to make predictions about a wide range of scenarios; while in the other, people have a large "bag of tricks", idiosyncratic heuristics that tend to work well in practice. In this paper, we suggest that rather than being in opposition to one another, these two ideas complement each other. We argue that people's capacity for mental simulation may support their ability to learn new cue-based heuristics, and demonstrate this phenomenon in two experiments. However, our results also indicate that participants are far less likely to learn a heuristic when there is no logical or explicitly conveyed relationship between the cue and the relevant outcome. Furthermore, simulation—while a potentially useful tool—is no substitute for real world experience.

Keywords: mental simulation, heuristics, physical reasoning

Introduction

The world is a complex place, yet people are able to navigate it effortlessly. How is the mind able to do so much? One answer is that the mind builds rich, generative models of the world (Tenenbaum, Kemp, Griffiths, & Goodman, 2011), which it then uses to "mentally simulate" potential futures and make inferences about objects and scenes. Indeed, there is a vast literature on how mental simulation underlies our core reasoning and problem solving abilities, including spatial reasoning (Hegarty, 2004; Shepard & Metzler, 1971), physical scene understanding (Battaglia, Hamrick, & Tenenbaum, 2013; Smith & Vul, 2013), counterfactual reasoning (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2014), and language comprehension (Bergen, Lindsay, Matlock, & Narayanan, 2007; Matlock, 2004). Yet, despite the power and flexibility of mental simulation, there is a cost associated with its use: running simulations and evaluating their results takes time and resources. An alternative is to rely instead on simple heuristics that usually point to a good answer (Gigerenzer & Todd, 1999). But, where do such heuristics come from in the first place?

Previous research has explored the notion of "learning by thinking" (Lombrozo, in press), demonstrating that people have the ability to learn new knowledge or re-represent old knowledge through internal processes such as simulation. For example, Hamrick, Battaglia, Griffiths, and Tenenbaum (2016) showed how people can use their mental simulations to learn about unobservable properties of the world such as the mass of objects; Khemlani, Mackiewicz, Bucciarelli, and Johnson-Laird (2013) illustrated how mental simulations can give rise to algorithmic problem-solving procedures; and Schwartz and Black (1996) demonstrated that people can learn simple rules about a physical system on the basis of mental simulation. Thus, it is clear that people *can* acquire new knowledge or heuristics from mental simulation; but, under what circumstances will they do so?

In this paper, we propose that generative models can bootstrap the discovery of heuristics for novel tasks, but that people's prior biases strongly influence how likely they are to discover such heuristics. We pose three key questions regarding this claim. First, to what extent are people able to learn new information from their mental simulations? Second, to what extent do people use this information to construct new heuristics? And third, is mental simulation as reliable as real-world experience in learning such heuristics?

To determine how people learn heuristics from mental simulation, we designed and ran two experiments adapted from Hamrick, Smith, Griffiths, and Vul (2015) in which participants predict whether or not a ball would go through a hole based on its initial trajectory (see Figure 1). Importantly, we also manipulated an environmental cue—the color of the box containing the ball—that perfectly predicted the correct response. In the first experiment, we primed participants with the knowledge that a simple rule existed (but did not tell them the rule itself); in the second, we primed them with either weak expectations or no expectations, and then allowed them to do the task and discover the rule independently. Our results show that people are capable of crystallizing new rules solely on the basis of their mental simulations, though they are significantly less likely to do so if they are not already entertaining the hypothesis that a rule exists. Moreover, we show that mental simulation, while an avenue for learning such rules, is no substitute for real world experience.

Experiment 1: Learning about known cues

In our first experiment, we asked to what extent people are able to learn heuristics from mental simulation when they are aware such a heuristic might exist. The heuristic took the form of an associative cue (see **Stimuli**) that perfectly predicted the correct response and that did not require mental simulation.

Methods

Participants We recruited 119 participants on Amazon's Mechanical Turk using the psiTurk experimental framework (Gureckis et al., 2015). Participants were paid $1.50 for roughly 14 minutes of work. We excluded 9 participants who did not finish the experiment and 8 participants who answered incorrectly on more than one catch trial. This left a total of 102 participants in our analysis.

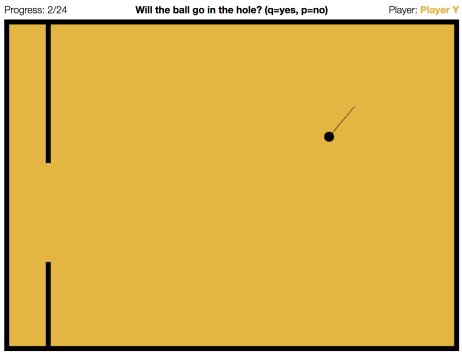

Figure 1: Example of a *medium* trial.

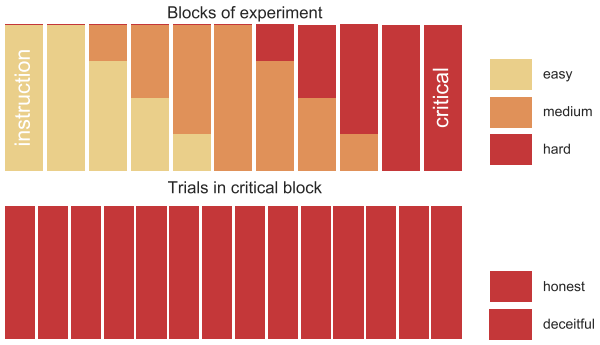

Figure 2: Trial structure. The experiment begins with a block of eight "instruction" trials, shown with feedback regardless of condition. No cue is present. This is followed by nine twelve-trial "standard" blocks of increasing difficulty. Fifty-four unique stimuli are each shown twice (in separate blocks), once with an *honest* cue and once with a *neutral* cue. Feedback is displayed on all or no trials depending on condition. At the end of each standard block, all participants saw their accuracy from the preceding block and responded to the *cue quiz* (see text). The final, "critical" block contains fourteen trials, shown without feedback. Trials with *deceitful* cues are interspersed to minimize the chance of participants noticing the change in cue reliability.

Design We used a $3 \times 3 \times 2$ mixed design. We manipulated two within-subject variables, CUE and DIFFICULTY. CUE could take on three values: *honest* (the cue perfectly predicts the correct response), *neutral* (the cue contains no information), and *deceitful* (the cue predicts the incorrect response). DIFFICULTY could take on three values as well: *easy*, *medium*, and *hard* (see **Stimuli**). We manipulated one between-subjects variable, FEEDBACK, which determined whether people were allowed to see the full path of the ball (and thus the correct answer) after making a judgment.

Stimuli The stimuli were animations of a ball moving at a 400px/s in a box with dimensions 900×650px. As the ball moved, it traced a gray line to reduce uncertainty about its direction. The initial stimulus presentation consisted of the ball moving for 0.2 seconds, after which the ball would freeze, remaining on screen along with its trace. The feedback animation picked up where the initial stimulus presentation left off, and showed the ball bouncing some number of times and then either (1) passing through the hole (a *hit*); or (2) bouncing off the central wall (a *miss*). The properties of a stimulus depended on the trial's difficulty. *Easy* stimuli had one bounce, a path length of 560px, and a hole size of 300px. *Medium* stimuli had one or two bounces, a path length of 880px, and a hole size of 200px. Finally, *hard* stimuli had two bounces, a path length of 1280px, and a hole size of 100px. The color of the background could be blue, green, or yellow depending on both the correct response and the value of CUE for that trial. For each participant, the three colors were mapped to *hit*, *miss*, and *neutral* (this mapping was counterbalanced). Thus, on an *honest* trial, the background would take the *hit* color if the ball would go through the hole, and the *miss* color, otherwise. This mapping was reversed for *deceitful* trials. Finally, on a *neutral* trial, the background was always the *neutral* color.

Procedure Participants were first given instructions in which the task was described. We specifically informed participants that they would observe three people playing a game on three different courts: "Player B" was playing on a blue court, "Player G" was playing on a green court, and "Player Y" was playing on a yellow court. We additionally told participants that one of the players was playing a game in which they were trying to get the ball in the hole, one was playing a game in which they were trying to avoid the hole, and one was playing a game in which they didn't care whether or not it went in. This backstory was designed to increase participants' subjective prior probability of and attention to the hypothesis that the background color was predictive of the correct response. Crucially, however, the backstory only motivated the existence of such a predictive relationship; it does not indicate its direction.

On each trial, participants were shown the scene, including the initial position of the ball and the location of the hole. Participants pressed 'space' to begin the trial, after which an animation of the initial stimulus began. Participants were then asked, "will the ball go in the hole?", and were instructed to press 'q' if they thought it would and 'p' otherwise. Participants in the *feedback* condition then saw "Correct!" or "Incorrect" as well as an animation showing the full remaining trajectory of the ball.

The structure of the experiment is shown in Figure 2. The early trials were *easy* so that participants in the *no feedback* condition had the chance to learn the cue when their simulation-based judgments were more reliable. The later trials were *hard* so they would be discriminative of participants' strategies: participants using simulation should perform poorly, while participants using the cue should be insensitive to trial difficulty. To assess declarative knowledge of the cue, we asked participants three multiple choice questions (the *cue quiz*) after each standard block: "Which player

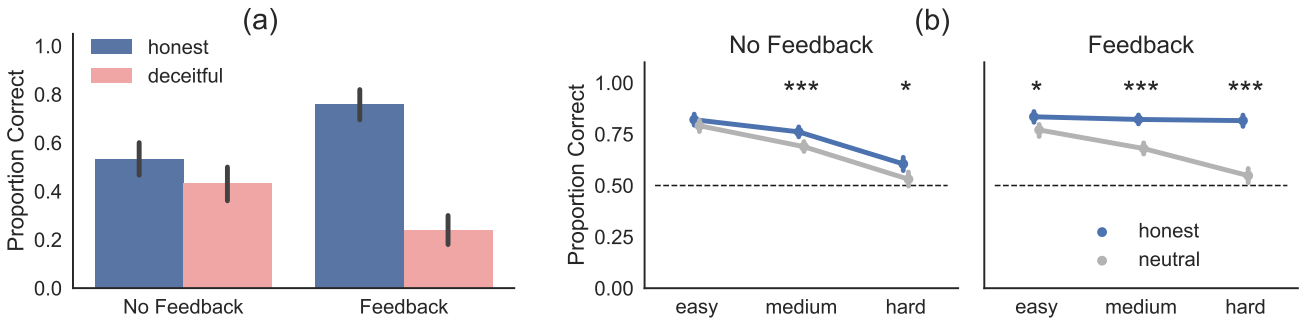

Figure 3: Accuracy on critical and standard trials in Experiment 1. Error bars in all figures denote 95% confidence intervals by bootstrapping. (a) Critical trials were displayed either with an honest cue or with a deceitful cue. Participants performed better when the cue is honest, suggesting that they were relying on the cue rather than using simulation. (b) On standard trials, participants tended to respond more accurately on honest-cue trials than on neutral-cue trials.

is trying to get the ball **into** the hole?", "Which player is trying to **avoid** the hole?", and "How confident are you in your response to the previous two questions?".

The last (critical) block was designed to provide evidence that participants were using the cue, as those using the cue should answer incorrectly on trials with the *deceitful* cue. The stimuli used for the first four *honest* trials and the *deceitful* trials were counterbalanced, and we excluded all *honest* trials after the first *deceitful* trial from analysis.

Results

All analyses were planned unless specifically stated otherwise, and all contrasts are adjusted for multiple comparisons. All data and analysis code can be viewed at https://osf.io/ut3xp/

Hypotheses Based on our experimental design, we hypothesized the following: (1) Participants in both the *no feedback* and *feedback* conditions will learn the cue, as determined by their responses to the cue quizzes and based on their responses on the critical trials. (2) Participants in both the *no feedback* and *feedback* conditions will use their knowledge of the cue to respond more accurately in the task. (3) Participants in the *feedback* condition will be more likely to learn and use the cue than participants in the *no feedback* condition.

Cue quizzes To gauge whether a participant had successfully learned the cue after seeing all standard trials, we restricted our analysis to the final cue quiz. We conducted three one-tailed proportion tests comparing the proportion of participants in each condition who answered both questions correctly on the quiz, with a chance probability of $\frac{1}{6}$. We found that 56% of participants in the *no feedback* condition ($\chi^2(1) = 54.465,\ p < 0.001$) and 68% of participants in the *feedback* condition ($\chi^2(1) = 91.204,\ p < 0.001$) correctly identified the cue. These results suggest that participants were able to use their simulations to learn about the cue, confirming our first hypothesis.

Critical trials According to our first hypothesis, we anticipated that participants who learned and used the cue strategy would fail on the four critical trials in which the cue was misinformative. We constructed a logistic regression model over accuracy on critical trials with factors for FEEDBACK and CUE. The results suggest that people in both the *feedback* and *no feedback* conditions were more likely to answer incorrectly on *deceitful* trials than on *honest* trials (Figure 3). Specifically, we found a significant main effect of CUE, with participants responding more accurately on *honest* trials than on *deceitful* trials ($\chi^2(1) = 4.252,\ p < 0.05$). We also found a significant main effect of FEEDBACK ($\chi^2(1) = 17.124,\ p < 0.001$), as well as an interaction between FEEDBACK and CUE ($\chi^2(1) = 40.019,\ p < 0.001$). In both feedback conditions people were more likely to answer incorrectly on *deceitful* trials than on *honest* trials, though this difference was only marginally significant in the *no feedback* condition (for *feedback*, LLR $= -2.31 \pm 0.23,\ z = -9.85,\ p < 0.001$; for *no feedback*, LLR $= -0.41 \pm 0.20,\ z = -2.06,\ p = 0.08$; where LLR is the log likelihood ratio).

The weak effect of cue honesty for the *no feedback* condition could be due to either an inability to identify the cue, or an inability to use knowledge of the cue to make predictions. To test these explanations, we conducted a post-hoc analysis identical to that above but restricting the data to those participants that passed the quiz. We found highly significant effects of CUE ($\chi^2(1) = 10.862,\ p < 0.001$), FEEDBACK ($\chi^2(1) = 30.657,\ p < 0.001$) ($\chi^2(1) = 17.124,\ p < 0.001$), and the interaction between FEEDBACK and CUE ($\chi^2(1) = 59.828,\ p < 0.001$). Contrasts revealed a significant effect of honesty in both the *feedback* (LLR $= -4.33 \pm 0.40,\ z = -10.85,\ p < 0.001$) and *no feedback* (LLR $= -0.88 \pm 0.27,\ z = -3.26,\ p < 0.01$) conditions. These results suggest that participants in the *no feedback* condition who identified the cue were also able to use it to make predictions, but not as well as those in the *feedback* condition.

Standard trials We also looked at the accuracy across trials during the main part of the experiment. We constructed a logistic regression model over accuracy with factors for FEEDBACK, DIFFICULTY, and CUE. The results are

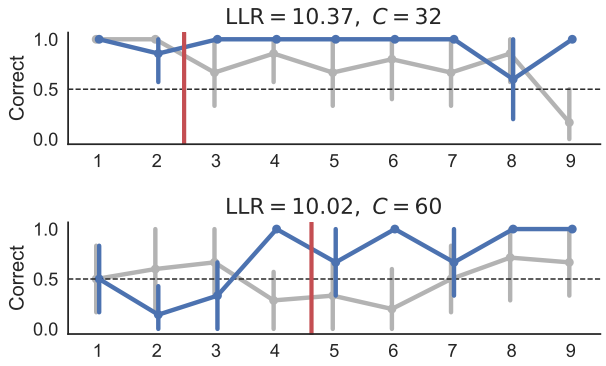

Figure 4: Example cue learners. Each subplot shows a different participant in the *no feedback* condition of Experiment 1 who was identified as learning the cue by our model. The blue lines indicate average accuracy on each block when using the honest cue, while the gray lines correspond to the neutral cue. The vertical red lines indicate the trial, C, when our model inferred they switched from using simulation to using the cue. The title of each subplot displays the log likelihood ratio of the *change* model to the *no change* model, as well as the trial when they changed strategies.

shown in Figure 3. We found a main effect of difficulty ($\chi^2(2) = 122.679,\ p < 0.001$), as well as a three-way interaction between FEEDBACK, DIFFICULTY, and CUE ($\chi^2(2) = 12.363,\ p < 0.01$).

We investigated differences in accuracy within feedback conditions and cue types and found that, overall, people were more accurate on *honest* trials when they had feedback than when they did not have feedback (LLR $= -0.51 \pm 0.07,\ z = -7.40,\ p < 0.001$). This supports our third hypothesis that real data is more reliable than simulated data. We did not detect a difference between feedback conditions on *neutral* trials, however, indicating that feedback did not affect people's accuracy when using simulation (LLR $= 0.03 \pm 0.06,\ z = 0.50,\ p = 1.00$). We also found that people were more accurate when the honest cue was present than when the neutral cue was present, both in the *feedback* condition (LLR $= -0.83 \pm 0.07,\ z = -12.24,\ p < 0.001$) and the *no feedback* condition (LLR $= -0.29 \pm 0.06,\ z = -4.57,\ p < 0.001$).

Modeling individual differences in cue learning

While the group-level effects in the previous sections confirmed our first and third hypotheses, we wanted to additionally investigate the individual behavior of participants who learned the cue. To this effect, we constructed a simple Markov model that allowed us to identify who actually used the cue and who did not.

Model For each participant, we defined a Markov model with observed states J_t representing the participants' judgment on trial t. For each strategy, we defined a probability of answering correctly. For the simulation probability, we fit $p_{\text{sim}}^{(\text{easy})}$, $p_{\text{sim}}^{(\text{med})}$, and $p_{\text{sim}}^{(\text{hard})}$ empirically based on the participant's average accuracy on trials without the cue for each level of difficulty. For the cue probability, we set $p_{\text{cue}} = 0.95$ to reflect a high probability of answering correctly, but not perfectly. Finally, we introduced a variable $C \in \{1, \ldots, T\}$ which indicated the "change point" at which participants switched from using simulation to using the cue heuristic.

The probability of a participant's judgment was then:

$$p(J_t = 1 \mid C) = \begin{cases} p_{\text{sim}}^{(d_t)} & t \leq C, \\ p_{\text{cue}} & t > C, \end{cases}$$

where d_t is difficulty of trial t. So, the probability of all responses was $\max_C p(\mathbf{J}_{1:T} \mid C) = \max_C \prod_{t=1}^{T} p(J_t \mid C)$, which we will refer to as the *change* model. We fit C in the change model to each participant separately.

We additionally computed the likelihood of participants' responses under a *no change* model, in which we computed $p(\mathbf{J}_{1:T} \mid C = \infty) = \prod_{t=1}^{T} p(J_t \mid C = \infty)$, where the infinite change point C indicates that the participant used the simulation strategy throughout the whole experiment.

Results To determine whether an individual participant learned the cue, we computed the log-likelihood ratio (LLR) between the *change* model and the null hypothesis (the *no change* model), and tested whether $2 \cdot \text{LLR}$ was significantly greater than zero under the χ^2 distribution, with a significance threshold of $p = 0.001$. Using this analysis, we found that 29 participants in the *feedback* condition switched to a cue-based strategy while 8 participants in the *no feedback* condition switched. To ensure these numbers were more than we would expect due to random chance, we additionally performed proportion tests with a probability of chance at 0.001 (corresponding to the significance threshold above). Both proportions were significantly different from chance (for feedback, $\chi^2(1) = 16204,\ p < 0.001$; for no feedback, $\chi^2(1) = 1068,\ p < 0.001$). The difference in proportions was also significant ($\chi^2(2) = 17272,\ p < 0.001$)

Figure 4 shows the two participants in the *no feedback* condition with the highest log-likelihood ratios, and illustrates the clear effect of the cue: on the *honest* trials, the participants have nearly perfect performance, while on the *neutral* trials, they are significantly worse.

We additionally looked at the overlap between those participants who correctly answered the cue quiz and those who were identified by our model. The results, shown in Table 1, indicate that those people who were identified by the model answered correctly on the quiz, but not necessarily the other way around. This suggests that, counter to our second hypothesis, not *everybody* who explicitly identifies the cue is able to apply that knowledge when performing the task.

Experiment 2: Discovering new heuristics

Based on the results of Experiment 1, it is clear that some people are able to use mental simulation to learn a cue-based heuristic—as long as they know that such a cue exists. In

Table 1: Number of participants identified by the quiz and/or model as having learned the cue in Experiments 1 and 2. FB = "Feedback", No FB = "No Feedback".

	Condition	**Neither**	**Quiz Only**	**Model Only**	**Both**
1	*No FB* (52)	20	24	3	5
	FB (50)	15	6	1	28
2A	*No FB* (48)	38	10	0	0
	FB (49)	31	14	1	3
2B	*No FB* (52)	41	11	0	0
	FB (49)	38	9	0	2

Experiment 2, we asked whether participants could discover and learn the heuristic without being given this information explicitly. By making two small alterations to the backstory presented in Experiment 1, we modulated the degree to which participants would expect the cue. Experiment 2A did not inform participants that the colors were predictive; it only associated the cue (color) with players. We hypothesized that this would allow participants to frame hypotheses about cue predictiveness in terms of more familiar concepts: one player might be more talented or have a different goal. Additionally, describing the colors in the instructions might increase their salience. In Experiment 2B, we did not verbally draw attention to the cue, nor did we provide any semantic meaning for the cue. Thus we expected participants would be even less likely to learn the cue, perhaps because they would not even consider the hypothesis that the colors are predictive.

Methods

Participants We recruited 224 participants on Amazon's Mechanical Turk using the psiTurk experimental framework (Gureckis et al., 2015). Participants were treated in accordance with UC Berkeley IRB standards and were paid $1.50 for fourteen minutes of work. We excluded 15 participants who did not finish the experiment and 11 participants who answered incorrectly on more than one catch trial. This left a total of 198 participants in our analysis.

Design and Procedure The design and procedure were identical to Experiment 1, with the following exceptions. In Experiment 2A we told participants that there were three different players, corresponding to three different colors, but not that they were playing different games. In Experiment 2B we gave participants a minimal backstory that made no reference to players or colors. In both experiments we administered the *cue quiz* once, at the end of the experiment.

Results

Experiment 2A We performed the same analyses as in Experiment 1, and found that 35% of people in the *feedback* condition were able to identify the cue in the quiz ($\chi^2(1) = 10.204,\ p < 0.001$). Without feedback, 21% of participants identified the cue, which was marginally significant ($\chi^2(1) = 1.680,\ p = 0.10$). We did not find an affect of CUE in the critical trials ($\chi^2(1) = 0.518,\ p = 0.47$), though there was a trend towards people being more accurate on *honest* trials. We found no significant effect of the cue on accuracy in standard trials either ($\chi^2(1) = 0.160,\ p = 0.69$).

The Markov model identified 4 people in the *feedback* condition ($\chi^2(1) = 243,\ p < 0.001$) and 0 in the *no feedback* condition as having adopted the cue strategy. Together, these results show that when people are primed with a cover story that makes the cue plausible, some of them will indeed learn the cue; however, the majority still will not.

Experiment 2B Whereas in Experiment 2A the cue was explained with a cover story about people playing a game, in Experiment 2B the cue was entirely unexplained. The results suggest that when the cue is unexplained, participants are unlikely to discover the informativeness of the cue. We again performed the same analyses as those in Experiment 1, and found that people were not significantly different from chance at identifying the cue in the survey, regardless of whether they saw feedback (21% of participants, $\chi^2(1) = 0.800,\ p = 0.19$) or not (22% of participants, $\chi^2(1) = 0.465,\ p = 0.25$). We also found no effect of CUE in the critical trials ($\chi^2(1) = 1.626,\ p = 0.20$), though as in Experiment 2A there was a trend toward people being more accurate on the *honest* trials. Again, we found no significant effect of the cue on accuracy in standard trials ($\chi^2(1) = 1.269,\ p = 0.26$).

The Markov model identified 2 people in the *feedback* condition ($\chi^2(1) = 43,\ p < 0.001$) and 0 in the *no feedback* condition as having adopted the cue strategy. These results suggest that when people are not already entertaining the hypothesis that a heuristic might exist, it is unlikely that they will spontaneously realize it.

Comparing Experiments Summary results of the three experiments are shown in Table 1 and Figure 5. We consistently find more evidence for cue-learning when feedback is given. However, our results suggest that an unexplained cue that has no intuitive relationship with the outcome is quite difficult to learn, even when feedback is present.

Conclusion

In this work, we asked three questions: (1) are people able to learn about auxiliary properties in the world through mental simulation; (2) do they use their knowledge to make more accurate predictions; and (3) is mental simulation as reliable as real-world experience? In Experiment 1, we showed that (1) people can indeed learn a correlated cue through the use of mental simulation; and (2) people can sometimes apply such knowledge as a heuristic prediction strategy. However, (3) both discovery and application of the cue was weaker when people had to learn from only simulated data. We speculate on two potential explanations for the advantage of external over simulated data. First, simulations are noisy; thus, simulated data may not accurately reflect the world. In this study,

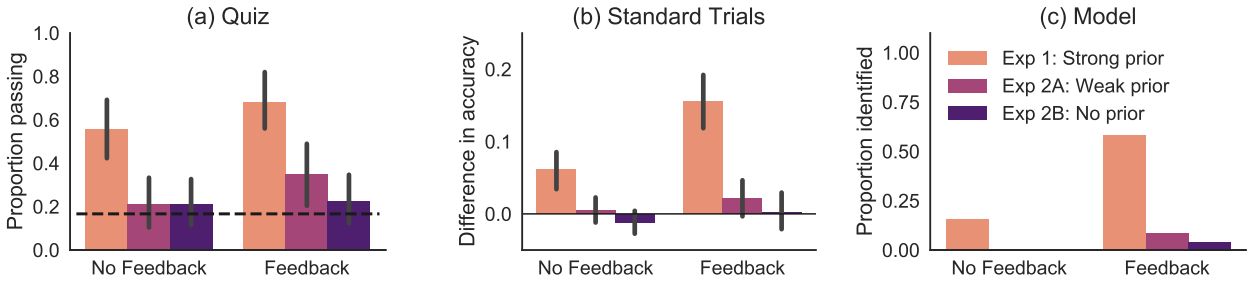

Figure 5: Comparing quiz, accuracy, and model results across experiments. (a) Participants who correctly identified the cue during the cue quiz. The dashed line indicates chance performance. (b) The difference in accuracy on the *honest* trials versus the *neutral* trials. Positive values indicate that participants were more accurate on *honest* trials. (c) The proportion of participants identified as having learned the cue by the Markov model. There are no error bars due to the particulars of this analysis; all non-zero proportions are significantly different from chance.

the cue was perfectly predictive; however, if one predicted incorrectly on 25% of trials, the cue would only 75% predictive. Furthermore, if people are aware that their simulations are error-prone, they may place less faith in the simulated data and any patterns therein. Second, simulations are costly, and it is possible that increased attentional and working memory load may have decreased participants' ability to simultaneously perform the task *and* pick up on the cue.

In both experiments, there was considerable within-condition variance in cue learning and use. The alignment between accuracy on the quiz and the Markov model predictions (Table 1) suggests that this is partly due to individual differences. It appears that some participants learned and applied the cue, while others completely ignored the cue. This between-subject variance could be due to true individual differences: perhaps some people are better able to learn associative cues (in general or specifically from simulated data). Alternatively, these differences could be the result of a constant learning ability that is stochastic and only occasionally expressed. Similar to flashes of intuition that strike seemingly at random, identifying a pattern in simulated data may be a powerful but rare event in human cognition.

Together, our results suggest that mental simulation on its own is not sufficient for learning: prior expectations are hugely important. This result is consistent with the ideas behind the hypothesis of theory-based causal induction (Tenenbaum, Griffiths, & Kemp, 2006), which posits that inductive reasoning requires highly structured and systematic systems of causal knowledge. While it was possible for participants in our experiments to learn a new piece of causal knowledge (a heuristic), it was very difficult for them to do so if the cue did not easily fit into an existing causal framework. Thus, we suggest that while mental simulation can be a powerful tool for re-representing knowledge, it does not operate in a vacuum, and must work in tandem with other cognitive processes to fully realize its potential.

Acknowledgements This work was supported by grant number ONR MURI N00014-13-1-0341.

References

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Bergen, B. K., Lindsay, S., Matlock, T., & Narayanan, S. (2007). Spatial and linguistic aspects of visual imagery in sentence comprehension. *Cognitive Science*, *31*(5), 733–64.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2014). From Counterfactual Simulation to Causal Judgment. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society.*

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. Oxford University Press, USA.

Gureckis, T. M., Martin, J. B., McDonnell, J. V., Alexander, R. S., Markant, D. B., Coenen, A., et al. (2015). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavioral Research Methods*, 2–16.

Hamrick, J. B., Battaglia, P. W., Griffiths, T. L., & Tenenbaum, J. B. (2016). Inferring mass in complex scenes by mental simulation. *Cognition*, *157*, 61 - 76.

Hamrick, J. B., Smith, K. A., Griffiths, T. L., & Vul, E. (2015). Think again? The amount of mental simulation tracks uncertainty in the outcome. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society.*

Hegarty, M. (2004). Mechanical reasoning by mental simulation. *Trends in Cognitive Sciences*, *8*(6), 280–285.

Khemlani, S. S., Mackiewicz, R., Bucciarelli, M., & Johnson-Laird, P. N. (2013). Kinematic mental simulations in abduction and deduction. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(42), 16766–71.

Lombrozo, T. (in press). "Learning by thinking" in science and everyday life. In P. Godfrey-Smith & A. Levy (Eds.), *The Scientific Imagination.* Oxford University Press.

Matlock, T. (2004). Fictive motion as cognitive simulation. *Memory & Cognition*, *32*(8), 1389–1400.

Schwartz, D. L., & Black, J. B. (1996). Shuttling between depictive models and abstract rules: Induction and fallback. *Cognitive Science*, *20*(4), 457–497.

Shepard, R. N., & Metzler, J. (1971). Mental Rotation of Three-Dimensional Objects. *Science*, *171*(3972), 701–703.

Smith, K. A., & Vul, E. (2013). Sources of uncertainty in intuitive physics. *Topics in Cognitive Science*, *5*(1), 185–199.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*(7), 309–318.

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.