Object Discovery and Inverse Physical Reasoning

Christopher D. Carroll (cdcarroll@gmail.com)
Department of Psychology, Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213

Charles Kemp (ckemp@cmu.edu)
Department of Psychology, Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213

Abstract

Unobserved objects are typically discovered by making backward inferences from effects to causes. The inverse reasoning account proposes that inferences of this kind are carried out by postulating unobserved causes that best support the forward inference from causes to effects. We evaluated the inverse reasoning account by asking people to reason about hidden attractors and repellers that caused an observed particle to move about an arena. We found that people often evaluated specific hypotheses in a manner consistent with the inverse reasoning account but that hypothesis discovery involved processes that were inconsistent with inverse reasoning.

Keywords: object discovery; inverse reasoning; inverse problem; Bayesian inference; physical reasoning

Introduction

Inferences about unobserved objects are common in both scientific and everyday reasoning. Scientists originally postulated the existence of the planet Neptune to explain perturbations in the orbit of Uranus. Similarly, a jilted lover may postulate the existence of a romantic competitor in order to explain the behavior of his or her partner. This paper describes an experimental study of object discovery that is loosely inspired by the discovery of Neptune. Participants observed particles that moved along paths such as the one in Figure 1 and attempted to infer the unobserved attractors and repellers responsible for the particle's motion.

Object discovery typically involves reasoning from effects (e.g., an observed motion) to causes (e.g., an unobserved attractor). Here we refer to inferences from causes to effects as *forward inferences* and inferences from effects to causes as *backward inferences*. We explore the hypothesis that forward and backward reasoning are tightly coupled, and that backward inferences are made by postulating unobserved causes that best support the forward inference from causes to effects. We refer to this approach as *inverse reasoning* because it achieves backward reasoning by inverting the process of forward reasoning.

One natural way to formalize the inverse reasoning approach makes use of Bayesian inference, which specifies the normative relationship between backward and forward reasoning. Specifically, given some observations D and a hypothesis H about the existence and properties of the unobserved causes, Bayes' theorem requires that

$$P(H|D) \propto P(D|H)P(H). \qquad (1)$$

Backward and forward reasoning are captured by the posterior $P(H|D)$ and likelihood $P(D|H)$, respectively. Bayes' theorem therefore suggests that backward reasoning should be carried out by combining the forward inferences specified by the likelihood with judgments of plausibility specified by the prior $P(H)$. In our setting, this approach suggests that a configuration of unobserved attractors and repellers is a good explanation for a particle's motion to the extent that (1) the configuration predicts the particle's motion and (2) the configuration is relatively parsimonious. Inverse reasoning implies that backward inferences will be consistent with forward inferences, but does not imply that backward inferences will always be accurate. Studies of physical reasoning have documented situations where people's forward inferences deviate from the predictions of classical mechanics (e.g., Clement, 1982; McCloskey, 1983), and faulty forward inferences could produce faulty backward inferences through inverse reasoning.

The inverse reasoning approach has a mixed record as an account of human reasoning. On one hand, the approach has been successfully used to develop models of causal reasoning (e.g., Griffiths & Tenenbaum, 2005), perception (e.g., Yuille & Kersten, 2006), sensorimotor control (e.g., Kording & Wolpert, 2006) and social reasoning (e.g., Baker,

Figure 1: This sequence of bird's-eye-view snapshots shows a particle's motion over time. The particle in this "wall-motion" scene moved along a diagonal path until it reached the top wall. It then continued along the wall.

Saxe, & Tenenbaum, 2009). On the other hand, psychologists have documented several respects in which backward inferences seem inconsistent with inverse reasoning (e.g., Kahneman, Slovic, & Tversky, 1982; Fernbach, Darlow, & Sloman, 2011). People often, for example, erroneously ignore or underutilize the prior when estimating the posterior (Bar-Hillel, 1980).

Based on these findings it is not clear whether the object discovery task considered in this paper should produce results that are consistent with inverse reasoning. Because physical reasoning is a core aspect of cognition that is present early in development (Spelke, Breinlinger, Macomber, & Jacobson, 1992), one might expect that backward physical reasoning will tend be consistent with normative inverse reasoning. Previous studies of physical reasoning provide some evidence for this claim. For example, Sanborn, Mansinghka, and Griffiths (2009) found that backward inferences about the relative masses of colliding objects were consistent with a Bayesian account of inverse reasoning. Object discovery, however, appears to be more challenging than the tasks considered by previous studies of backward physical reasoning. Inferring hidden properties of observed objects (e.g., the mass of a colliding object) is a relatively well-constrained problem, but object discovery is a more open-ended problem that involves inferring the existence and number of the hidden objects, the locations of those objects, and the properties of those objects. To preview our results, we found that when participants evaluated specific hypotheses about the locations and properties of the hidden objects, their inferences were broadly consistent with inverse reasoning. When asked to generate their own explanations, however, many participants gave responses that were incompatible with the inverse reasoning account.

Experimental overview

To explore the problem of object discovery we conducted an experiment where participants reasoned about "attractors" and "repellers" that controlled the movements of some observed "particles." The attractors and repellers were unobserved, and participants attempted to infer their locations given the observed particle motions.

There were three experimental phases: the discovery, prediction, and evaluation phases. In the discovery phase, participants observed the motion of a particle and were asked to infer the locations of hidden attractors and repellers. In the prediction phase, participants were given the locations of one or more attractors or repellers and were asked to predict the trajectory that a particle would follow. In the evaluation phase, participants were given two possible explanations of a particle motion and were asked to decide which explanation was better. Note that the discovery and evaluation phase both assessed backward reasoning and that the prediction phase assessed forward reasoning.

The simplest possible observed trajectory is a straight line, and the obvious explanation for this trajectory is that the particle is either moving towards an attractor or moving away from a repeller. The particle motions presented in the discovery phase (Figure 2.i) include some of the next simplest cases. Each motion can be explained in at least two ways. First, there is a parsimonious explanation that invokes a relatively small number of stationary attractors and repellers. For example, the "wall-motion" scene (Figure 1 and Figure 2.a.i) can be explained by assuming a single repeller (see the first row of Figure 2.ii). Second, each explanation had a less parsimonious explanation where the particle always moved directly towards an attractor or directly away from a repeller, but where the attractors and repellers spontaneously appeared, disappeared, or moved. The second row of Figure 2.ii shows a less parsimonious explanation of the wall-motion scene.

Our primary goal is to explore whether participants generate the parsimonious explanations during the discovery and evaluation phases. If participants agree that the parsimonious explanations are in fact parsimonious and valid, then the inverse reasoning account predicts that these explanations should be generated during the discovery phase and rated favorably during the evaluation phase. If participants fail to generate these explanations in the discovery phase but tend to prefer them in the evaluation phase, this result would be inconsistent with the inverse reasoning view.

The prediction phase asked participants to generate particle trajectories for several kinds of configurations. Each configuration can be viewed as an explanation (plausible or implausible) of a motion observed during the discovery phase. Some of the prediction trials presented participants with their own explanations from the discovery phase. For our purposes, however, the most important prediction trials are those that presented participants with the parsimonious explanations for the three motions in Figure 2.i. Including these trials allowed us to assess whether participants agreed that the parsimonious explanations could in fact explain the observed motions – if not, it would be unsurprising if these explanations were rarely chosen during the discovery phase.

Method

Participants

Thirty undergraduates at Carnegie Mellon University participated for course credit.

Materials and Procedure

Participants were asked to imagine themselves working for a scientist who studies "attractors" and "repellers." The instructions explained that the participants would view scenes where "particles" moved within a rectangular arena. Participants learned that the particle motions were caused by attractors and repellers located outside the arena.

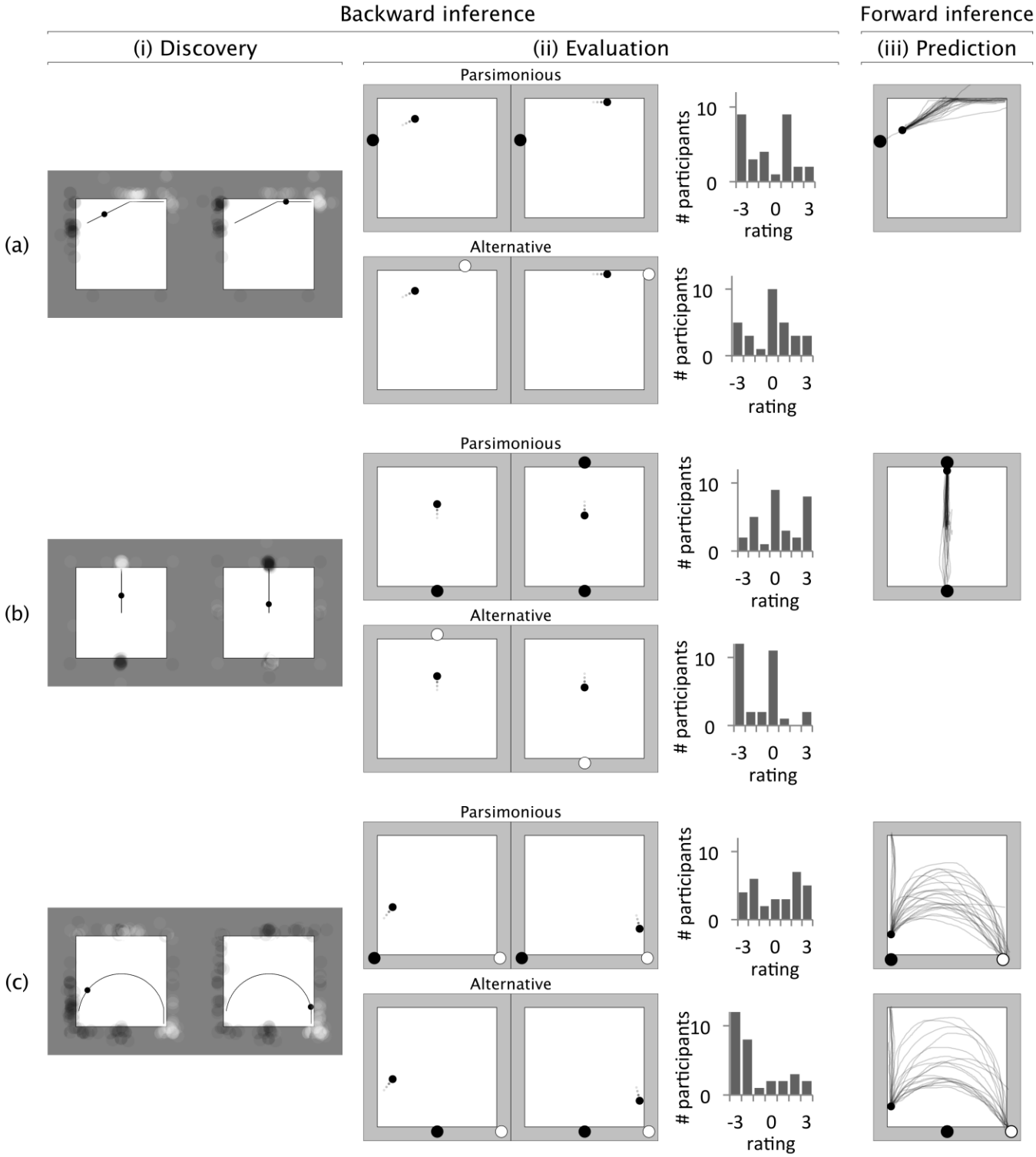

Figure 2: Experimental method and results for the (a) wall-motion, (b) center-return, and (c) curved-motion scenes. (i) The discovery phase. The paths illustrate the motion of the particle, and the circles illustrate the location of the particle in the first and second response pictures. (The particle in the center-return scene moved from its initial position to the top wall, paused, and then returned to the center.) The surrounding area is a heatmap. Areas where attractors were often placed are shown as brighter areas and areas where repellers were often placed are shown as darker areas. (ii) The evaluation phase. The pictures at left show the experimenter-provided explanations. Repellers and attractors are shown as large black and large white circles, respectively. The histograms shows the preference ratings (-3 = strongly preferred own explanation; 3 = strongly preferred experimenter-provided explanation). (iii) The prediction phase. The paths in the figure represent the paths drawn by the participants. Trivial prediction trials (i.e., those involving a single attractor or repeller) are not shown.

Participants then viewed three scenes that demonstrated the properties of the attractors and repellers. Each scene was displayed as a sequence of bird's-eye-view snapshots showing the motion of the particle over time. The first two scenes showed that particles move towards attractors and away from repellers, depicted as green and red circular objects, respectively. The third scene showed that a particle placed between two attractors moved towards the closer one, and the instructions explained that distant attractors and repellers exert less force than close ones.

Discovery phase

Participants were asked to explain a number of scenes where the attractors and repellers were not visible. After completing a practice trial, participants generated explanations for the three scenes in Figure 2.i. Participants also generated explanations for 12 variants of the three primary scenes, but we do not discuss these results here because the variant scenes did not have analogues in the prediction and evaluation phases. In the wall-motion scene (Figure 2.a.i), the particle traveled along a diagonal until it reached the top wall of the arena. It then continued along the top wall of the arena. In the center-return scene (Figure 2.b.i), the particle moved from the center of the arena to the top wall, paused, and then returned to the center. In the curved-motion scene (Figure 2.c.i), the particle moved along a curved path from the lower-left corner of the arena to the lower-right corner of the arena.

Participants explained each particle motion by specifying where the attractors and repellers would have been at two different points in the particle's motion (see Figure 2.i). The instructions explained that the participants were being asked to report the locations of the attractors and repellers in two distinct response pictures because "there may be some situations where you think that something has changed." Responses were made using a computer interface that showed the two response pictures and a summary of the to-be-explained particle motion. Participants could place attractors and repellers by clicking on any location outside the arena. Participants could move or erase placed attractors and repellers. A "reuse" button located between the two response pictures copied the attractors and repellers in the first picture to the second picture.

Participants were allowed to provide up to three explanations for each scene. Each explanation was entered on a separate screen. Participants were allowed to provide written explanations to supplement the picture-based explanations, but few participants did so.

After providing the explanations, the participants rated each provided explanation on a scale ranging from 1 (very unlikely to be the true explanation) to 7 (very likely to be the true explanation). Participants were also asked to rate the likelihood that the true explanation was "fundamentally different" from the provided explanation(s).

Prediction phase

Participants were asked to predict the particle paths given the locations of the attractors and repellers. Figure 2.iii presents some of the prediction trials. The prediction pictures in Figure 2.a.iii, b.iii, and c.iii top correspond to the parsimonious explanations. There were other prediction trials that corresponded to less parsimonious explanations. Three other prediction trials presented each participant with the configurations that corresponded to his or her own explanations in the discovery phase.

Evaluation phase

In the evaluation phase, participants once again viewed the wall-motion, center-return, and curved-motion scenes. In explaining each scene, participants chose between their own explanations and the parsimonious explanation. The parsimonious explanations are shown as the first, third, and fifth rows in Figure 2.ii.

For each forced choice, the participant rated his or her preferred explanation as "much more," "more", or "slightly more" likely to be the true explanation than the competing explanation. Because participants occasionally generated the parsimonious explanations themselves, participants were sometimes presented with a choice between two identical explanations. For these situations, participants were provided with a "these explanations are identical" button. We coded responses on a scale ranging from -3 (own explanation "much more likely" to be the true explanation) to 3 (parsimonious or alternative explanation "much more likely" to be the true explanation). When a participant claimed that the explanations were identical, his or her preference was coded as 0.

Three other trials required the participants to choose between their own explanations and some less parsimonious explanations. These alternative explanations, shown in the second, fourth, and sixth rows of Figure 2.ii, required additional assumptions to explain the particle motion. These trials served to control for the task demand of asking the participants to choose between their own explanation and an experimenter-provided explanation. To further limit any task demands, all competing explanations were described as responses provided by other participants.

Results

The inverse reasoning account predicts that participants ought to generate the parsimonious explanations during the discovery phase and endorse them during the evaluation phase. In contrast, we found that participants rarely generated the parsimonious explanations during the discovery phase but often preferred them during the evaluation phase. We begin by documenting this general result and then provide more detailed descriptions of the results for the discovery and prediction phases.

Parsimonious explanations

A wall-motion explanation was coded as parsimonious when it invoked a single stationary attractor or repeller. A center-return explanation was coded as parsimonious when it invoked two balancing repellers above and below the arena or two balancing attractors to the left and right of the

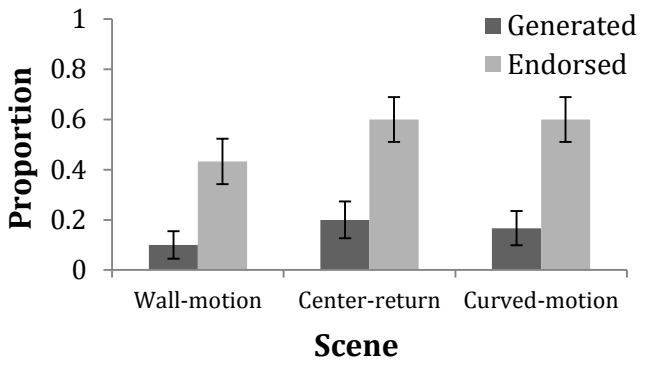

Figure 3: Proportions of the participants generating a parsimonious explanation in the discovery phase and endorsing the parsimonious explanation in the evaluation phase.

arena. A curved-motion explanation was coded as parsimonious when it invoked exactly two stationary attractors or repellers and did not invoke any moving, appearing, or disappearing attractors and repellers. Our coding criteria were intended to be conservative: note, for example, that any curved-motion explanation with two stationary objects was coded as parsimonious regardless of the locations of these objects.

Figure 3 shows that participants rarely generated parsimonious explanations in the discovery phase but often preferred them in the evaluation phase. The differences between the rates of generation and endorsement were significant for each scene (Fisher's exact test yields $p < .01$ in all cases). This finding cannot be attributed to task demands alone: as shown by the distribution of the preference ratings in Figure 2.ii, participants did not prefer non-parsimonious explanations (rows two, four, and six) to the same extent that they preferred the parsimonious explanations (rows one, three, and five).

Discovery

The difference between the results for the discovery and evaluation phases suggests that object discovery in our paradigm is not accurately characterized as inverse physical reasoning. Figure 2.i gives some sense of how participants were approaching the discovery task. Each plot in this column is a "heatmap:" locations where participants often placed attractors are shown as brighter areas and locations where repellers were often placed are shown as darker areas.

For the wall-motion trials, 14 participants posited one hidden object along the particle's diagonal trajectory and one hidden object along its horizontal trajectory, 5 participants posited a single attractor or repeller that moved, and 8 participants generated combinations or variations of those explanations. For the center-return trials, 12 participants posited appearing and disappearing attractors and repellers along each path of motion, 9 generated an explanation that involved balancing attractors or repellers but also invoked other attractors and repellers (e.g., had balancing repellers in the second response picture but

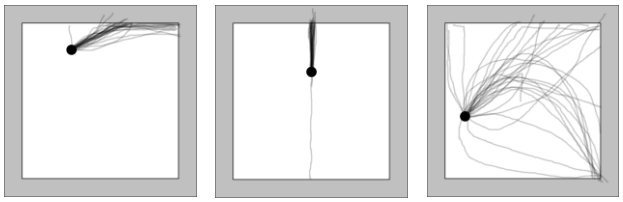

Figure 4: Particle motions predicted by participants given their own explanations for the wall-motion, center-return, and curved-motion scenes.

posited an attractor in the first response picture), and 5 participants generated other explanations. Responses to the curved-motion scene were more variable, and there was less agreement on the locations of the attractors and repellers (see Figure 2.i.c). Most of the participants posited multiple attractors and repellers that were simultaneously present. For example, 7 participants posited three or more stationary and constantly present attractors and repellers, and 12 participants posited two or more attractors or repellers that were simultaneously present at some point during the motion but were either non-stationary or not constantly present. The remaining non-parsimonious explanations most commonly posited a single attractor or repeller along the particle's path of motion in each response picture. Overall, then, responses to the discovery phase reveal a variety of strategies, but one consistent element is that many participants placed objects in line with a particle's instantaneous direction of motion.

Prediction

Figure 2.iii summarizes the responses on selected prediction trials. The critical question for present purposes is whether participants agreed that the parsimonious explanations would indeed account for the observed motions during the discovery trials. When provided with the parsimonious explanations, 15 of the 30, 24 of the 30, and 23 of the 30 participants predicted that the particle would approximately reproduce the particle motions from the discovery trials for the wall-motion, center-return, and curved-motion scenes respectively. Note that these counts are substantially higher than the number of participants who generated the parsimonious explanations during the discovery phase. The prediction data therefore provide further evidence that some participants failed to generate the parsimonious explanations during the discovery phase even though they considered these explanations to be valid.

Although participants did not always predict that the parsimonious explanations would produce the observed motion, their predictions were usually sensible given some additional assumptions. For example, various participants seemed to assume that friction would stop the particle when it hit the wall in the parsimonious wall-motion prediction trial, that momentum would carry the particle past the center in the parsimonious center-return prediction trial, and that the motion of the particle would be influenced *only* by

nearby repeller in the parsimonious curved-motion prediction trial.

Figure 4 shows the participants' predictions given their own explanations for the wall-motion, center-return, and curved-motion scenes. Some predicted motions diverged dramatically from the particle motion in the to-be-explained scene, and the discrepancies for the curved-motion scene were especially dramatic. These discrepancies should be interpreted cautiously, however, because the participants may have made different assumptions during the discovery and prediction phases. For example, it was natural to assume that the particle had an initial velocity in the first response picture of a trial in the discovery phase (the particle had already moved), but there was no reason to assume a particle velocity in the prediction phase. As a result, future studies are needed before concluding that participants sometimes generate explanations that are truly incompatible with the trajectories that they have observed.

Discussion

Our data support the conclusion that hypothesis evaluation is consistent with the inverse reasoning account but that hypothesis discovery is not. In some respects, the failure of the participants to discover the parsimonious explanations is quite surprising. The parsimonious explanations were straightforward, requiring the participant to posit at most two stationary attractors or repellers. It should have been possible for participants to discover the parsimonious explanations, and some of them indeed did so. In other respects, the failure of the participants to discover the parsimonious explanation makes sense. Even in the simple object discovery task presented in this paper, there are infinitely many explanations that might be considered. The inverse reasoning account is often unhelpful in these situations. Bayes' theorem admonishes the reasoner to consider all the possible explanations for the observations, but does not provide guidance when doing so is impossible.

Although generating the best explanation from an infinite class may be computationally challenging, evaluating the merits of a handful of selected hypotheses seems substantially easier. It is therefore not surprising that hypothesis evaluation was broadly – although perhaps not absolutely – consistent with the normative inverse reasoning account. The dissociation between discovery and evaluation is consistent with the view that people rely on non-Bayesian strategies to generate candidate explanations for evaluation, but are able to approximate Bayesian reasoning when deciding which of these candidates is best (Bonawitz and Griffiths, 2010).

Participants may have used several different kinds of strategies to generate candidate explanations during the discovery phase of our experiment. For example, an initial explanation might have been generated using the idea that objects often move directly towards attractors. If needed, this initial explanation might have been improved using search heuristics such as hill-climbing. The process of discovery might also rely on analogical reasoning—for example, many participants explained the curved-motion scene by placing a repeller at the focal point of the curve, and it is tempting to view this inference as a loose analogy to orbital motion. Like any other kind of creative behavior, object discovery is likely to be difficult to characterize in full detail. Future studies, however, can aim to characterize some of the psychological processes involved.

Acknowledgments

This work was supported in part by NSF grant CDI-0835797 and by the Pittsburgh Life Sciences Greenhouse Opportunity Fund.

References

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition, 113* , 329-349.

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

Bonawitz, E. B., & Griffiths, T. L. (2010). Deconfounding hypothesis generation and evaluation in Bayesian models. In S. Ohlsson, & R. Catrambone (Eds.), *Proceedings of hte 31st Annual Conference of the Cognitive Science Society* (pp. 2260-2265). Austin, TX: Cognitive Science Society.

Clement, J. (1982). Students' preconceptions in introductory mechanics. *American Journal of Physics, 50* , 66-71.

Fernbach, P. M., Darlow, D., & Sloman, S. A. (2010). Neglect of alternative causes in predictive but not diagnostic reasoning. *Psychological Science, 21 (3)* , 329-336.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51* , 334-384.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases.* Cambridge: Cambridge University Press.

Kording, K. P., & Wolpert, D. M. (2006). Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences, 10 (7)* , 319-326.

McCloskey, M. (1983). Intuitive physics. *Scientific American, 24* , 122-130.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review, 117 (4)* , 1144-1167.

Spelke, E. S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of knowledge. *Psychological Review, 99 (4)* , 605-632.

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences, 10 (7)* , 301-308.