Cognition 139 (2015) 130–153

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Evaluating the inverse reasoning account of object discovery [☆]

Christopher D. Carroll [*], Charles Kemp

Department of Psychology, Carnegie Mellon University, United States

ARTICLE INFO

Article history:
Received 19 December 2013
Revised 1 December 2014
Accepted 4 March 2015

Keywords:
Object discovery
Inverse reasoning
Bayesian inference
Physical reasoning

ABSTRACT

People routinely make inferences about unobserved objects. A hotel guest with welts on his arms, for example, will often worry about bed bugs. The discovery of unobserved objects almost always involves a *backward inference* from some observed effects (e.g., welts) to unobserved causes (e.g., bed bugs). The inverse reasoning account, which is typically formalized as Bayesian inference, posits that the strength of a backward inference is closely connected to the strength of the corresponding *forward inference* from the unobserved causes to the observed effects. We evaluated the inverse reasoning account of object discovery in three experiments where participants were asked to discover the unobserved "attractors" and "repellers" that controlled a "particle" moving within an arena. Experiments 1 and 2 showed that participants often failed to provide the best explanations for various particle motions, even when the best explanations were simple and when participants enthusiastically endorsed these explanations when presented with them. This failure demonstrates that object discovery is critically dependent on the processes that support hypothesis generation—processes that the inverse reasoning account does not explain. Experiment 3 demonstrated that people sometimes generate explanations that are invalid even according to their own forward inferences, suggesting that the psychological processes that support forward and backward inference are less intertwined than the inverse reasoning account suggests. The experimental findings support an alternative account of object discovery in which people rely on heuristics to generate possible explanations.

1. Introduction

Some of the most celebrated discoveries in the history of science involve inferences about unobserved objects. Nineteenth-century astronomers, for example, inferred Neptune's existence on the basis of mathematical calculations showing that an unseen planet would explain anomalies in the orbits of the known planets. Inferences about unobserved objects are also common, though less dramatic, in everyday reasoning. A hotel guest who awakes with welts on his arms will suspect bed bugs, and a person who has been jilted by a lover will worry about romantic competitors.

Because object discovery represents a particularly dramatic inductive leap, it often attracts the attention of thinkers who are interested in inductive inference. Philosophers of science, for example, view the task of understanding object discovery as one of the central problems of their field (e.g., Churchland & Hooker, 1985; van Fraassen, 1980), and specific episodes such as the discovery of Neptune have drawn repeated comment (e.g., Howson & Urbach, 1989/1996, pp. 121–122; Jaynes, 2003, pp. 133–139; Lipton, 2004, p. 89; Polya, 1954, pp. 130–132). Psychologists, however, have largely neglected

[☆] The preparation of this article was supported by NSF grant CDI-0835797, NIMH training grant T32MH019983 and the Pittsburgh Life Sciences Greenhouse Opportunity Fund. A preliminary report of this research was presented at the 34th Annual Conference of the Cognitive Science Society.

[*] Corresponding author at: Department of Psychology, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, United States.

E-mail addresses: cdcarroll@gmail.com (C.D. Carroll), ckemp@cmu.edu (C. Kemp).

http://dx.doi.org/10.1016/j.cognition.2015.03.003
0010-0277/© 2015 Elsevier B.V. All rights reserved.

the problem of object discovery. There are several studies that demonstrate that people—and even infants—discover unobserved objects in some circumstances (e.g., Csibra & Volein, 2008; Saxe, Tenenbaum, & Carey, 2005), but the general principles that support these inferences are not well understood.

In this paper, we evaluate one potential account of object discovery: the *inverse reasoning account*. Object discovery often involves reasoning from effects to causes (e.g., reasoning from orbital anomalies to an undiscovered planet), and these *backward inferences* can be contrasted with *forward inferences* that involve reasoning from causes to effects.[1] The inverse reasoning account posits that backward inferences are made by "inverting" the process of reasoning forward from causes to effects. As a result, the strength of a backward inference is closely related to the strength of the corresponding forward inference. The inverse reasoning account is usually formalized as Bayesian inference, which provides a normative framework for backward inference.

There are at least two reasons why inverse reasoning demands consideration as an account of object discovery. First, philosophers of science have proposed that object discovery—and scientific reasoning more generally—can be viewed as a form of Bayesian inference (e.g., Howson & Urbach, 1989/1996). Second, even though psychologists have not yet evaluated Bayesian inference as an account of object discovery, they have developed Bayesian models of cognition that characterize human inference in a wide variety of inferential tasks. The phenomena considered range from low-level processes that support perception (e.g., Ernst & Banks, 2002; Yuille & Kersten, 2006), motor planning (e.g., Kording & Wolpert, 2006), language processing (e.g., Chater & Manning, 2006), and semantic memory (e.g., Griffiths, Steyvers, & Tenenbaum, 2007) to higher-level processes that support inferences about object dynamics (e.g., Battaglia, Hamrick, & Tenenbaum, 2013; Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2012; Sanborn, Mansinghka, & Griffiths, 2013; Smith & Vul, 2013; Teglás et al., 2011), causation (e.g., Griffiths & Tenenbaum, 2005; Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008), property generalization (e.g., Kemp & Tenenbaum, 2009), and social agents (e.g., Baker, Saxe, & Tenenbaum, 2009). It is therefore worth exploring whether inverse reasoning can also explain how people discover unobserved objects.

To evaluate inverse reasoning as a psychological account of object discovery, we investigated inference in an experimental setting that was loosely inspired by the discovery of Neptune. Participants in our experiments observed particle motions such as the ones shown in Fig. 1 and attempted to discover the unobserved "attractors" and "repellers" that influenced the particles. In addition to completing this sort of *discovery trial*, our participants also completed *prediction* and *evaluation trials*: the prediction trials required them to predict the motion of a particle given an observed configuration of attractors and repellers, and the evaluation trials required them to choose between two possible explanations of a particle motion. The discovery and evaluation trials involve backward inferences and the prediction trials involve forward inferences. The inverse reasoning account predicts that a reasoner's inferences on these different tasks will be consistent, and our experiments were designed to test this prediction.

We expected that forward inferences in our experimental setting would be intuitive and straightforward because physical reasoning is a core aspect of cognition that is present early in development (Spelke, Breinlinger, Macomber, & Jacobson, 1992; Teglás et al., 2011) and that draws on a rich set of intuitions about physical causation (e.g., diSessa, 1993). Although people's intuitions about physical causation are not always accurate (e.g., Clement, 1982; diSessa, 1993; McCloskey, 1983; McCloskey, Caramazza, & Green, 1980), this does not present a problem for our experimental strategy. The inverse reasoning account, after all, does not claim that inferences on discovery, prediction, and evaluation tasks will always be accurate; it claims only that inferences on these tasks will be mutually consistent. We hoped that choosing an experimental setting in which forward inference is straightforward would allow us to focus on the consistency of these inferences.

Although we focus throughout on the inverse reasoning approach, an alternative tradition characterizes scientific discovery as a problem-solving task in which the scientist searches for a theory that can explain the observed data (e.g., Klahr & Dunbar, 1988; Langley, Simon, Bradshaw, & Zytkow, 1987; Simon, Langley, & Bradshaw, 1981). As part of this tradition, psychologists and computer scientists have developed computational models that recapitulate several historical examples of scientific discovery. For example, the DALTON system (Langley et al., 1987) has been used to model the discovery of the structure of substances involved in chemical reactions, and the GELL-MANN system (Fischer & Zytkow, 1992) has been used to model the discovery of subatomic particles such as quarks. We will return to the problem-solving approach in the general discussion, and will discuss how it relates to the inverse reasoning approach and the extent to which it is consistent with our data.

1.1. The inverse reasoning account

The inverse reasoning account can be formalized using Bayes' theorem, which establishes the normative relationship between forward and backward inferences in probabilistic settings. Given data d and a hypothesis h, Bayes' theorem states that

$$P(h|d) = \frac{P(d|h)P(h)}{P(d)}. \tag{1}$$

In the context of our task, the data d is the observed particle motion (the effect) and the hypothesis h represents a possible configuration of the unobserved attractors and repellers (the cause). Forward and backward inferences are captured by the likelihood $P(d|h)$ and posterior $P(h|d)$,

[1] We will use the term *backward inference* to refer to any inference from observed effects to unobserved causes, and will reserve the term *inverse reasoning* to refer to a specific approach to evaluating backward inferences. This distinction is not always made in other contexts: other researchers refer to the "inverse problem" and "inverse models" (e.g., Fienberg, 2006; Tarantola, 2006; Wolpert & Kawato, 1998) in contexts where we would refer only to *backward inference*.

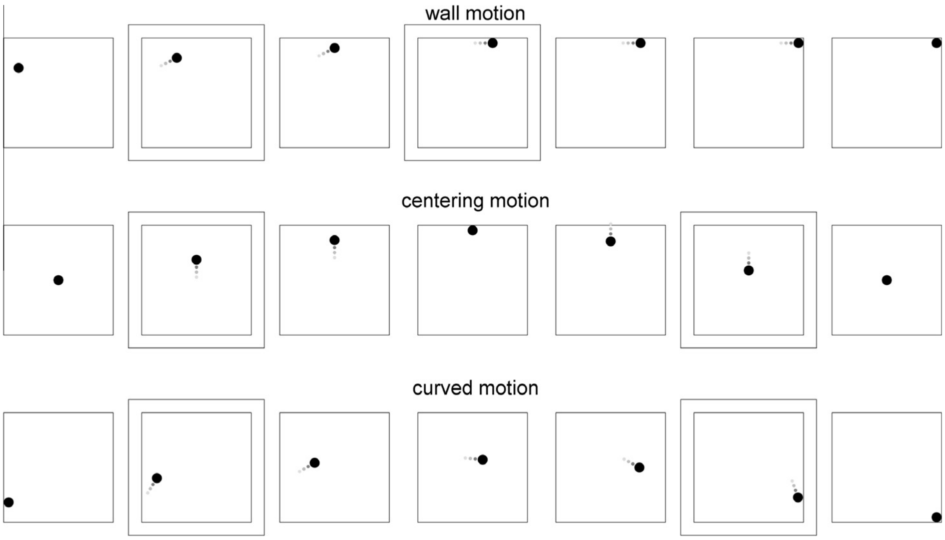

Fig. 1. Particle motions used to study backward inference. In each case, participants viewed a sequence of "snapshots" and were told that the particles were influenced by unobserved "attractors" and "repellers" located outside of the arena. The figure shows abbreviated snapshot sequences for the wall, centering, and curved motions from Experiment 1. Participants provided explanations of the motions by indicating where the attractors and repellers would have been in the outlined snapshots.

respectively, and the prior $P(h)$ captures the *a priori* plausibility of the hypothesis. We assume that the prior probability of a hypothesis is based on its simplicity (Lombrozo, 2007): for example, configurations with few attractors and repellers are more plausible *a priori* than configurations with many attractors and repellers. The denominator ensures that the posterior probabilities of the hypotheses sum to 1.0.

The exact commitments of the inverse reasoning account depend on whether it is formulated as a computational- or process-level account of object discovery (see Marr, 1982). We consider the commitments of each of these formulations in turn.

1.2. Inverse reasoning as a computational-level account

When formulated at the computational level, the inverse reasoning account proposes that people are able to identify the explanation that maximizes the posterior probability in Eq. (1) (see the upper half of Fig. 2). This formulation predicts which explanation the reasoner will provide, but does not make any process-level commitments about how people discover or identify this explanation.

How likely is it that people will be able to discover the most probable explanations in our object-discovery task? Different considerations suggest different answers. On the one hand, given the early emergence of physical reasoning in cognitive development and given the importance of physical reasoning in everyday inference (Spelke et al., 1992), people should be highly skilled at making inferences about physical events. This prediction has been tested in a task where participants infer the mass ratio of two objects by viewing scenes in which those objects collide (e.g., Gilden & Proffitt, 1989; Todd & Warren, 1982), and people's inferences on this task are largely consistent with the inverse reasoning account (Sanborn et al., 2013).

On the other hand, computational-level accounts predict human inferences more accurately on some tasks than on others (Marcus & Davis, 2013; see also Fernbach & Sloman, 2009), and it may be more difficult to discover the best explanation in our object-discovery task than in the mass-ratio task. After all, the object-discovery task involves simultaneously inferring the number, the locations, and the properties of multiple unobserved objects; the mass-ratio task only involves inferring a single unobserved property (the mass ratio). In tasks such as object discovery, it can be challenging to decide which explanations to consider in the first place, and people sometimes fail to consider relevant explanations (e.g., Bonawitz & Griffiths, 2010). Further evidence for the importance of this decision comes from the finding that people frequently overestimate the probability of a focal explanation. This overestimation is commonly attributed to a failure to consider alternative explanations (e.g., Tversky & Koehler, 1994; see also Koriat, Lichtenstein, & Fischoff, 1980; Thomas, Dougherty, Sprenger, & Harbison, 2008).

In evaluating the computational-level inverse reasoning account, we were particularly interested in comparing inferences on the discovery and evaluation trials. Recall that the evaluation trials required participants to choose between two possible explanations of a particle motion. The computational-level account in Fig. 2 is applied to these trials by restricting the hypothesis space H to contain only the two candidate explanations provided. Because the computational-level account predicts that participants will identify the explanation with the highest possible posterior probability on the discovery trials, it predicts that participants will prefer the explanation identified on the discovery task to any other explanation presented during the evaluation trials. To test this prediction, we presented participants with evaluation trials that asked them to

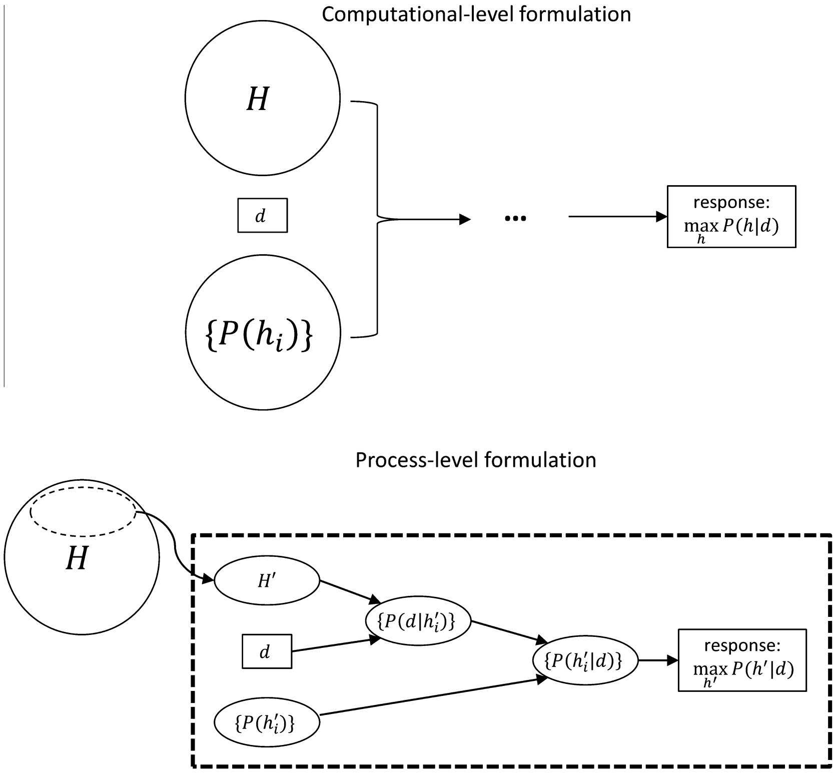

Fig. 2. Two ways in which the inverse reasoning account can be applied to object discovery. The computational-level formulation predicts that people can identify the explanation with the maximum posterior probability, but does not entail any commitments about how this explanation is identified. The process-level formulation suggests that people use Eq. (1) to evaluate a subset of the possible explanations by performing inverse reasoning over that subset. The dashed rectangle outlines the scope of the process-level inverse-reasoning account. d = the observed data; H = the hypothesis space of possible explanations; H' = a subset of the full hypothesis space; $\{P(h)\}$ = the prior probabilities of each hypothesis in H; $\{P(h')\}$, $\{P(d|h')\}$, and $\{P(h'|d)\}$ = the prior probabilities, likelihoods, and estimated posterior probabilities of each hypothesis in H'.

choose between their own explanations from the discovery trials and various alternative explanations. To the extent that discovering the explanation with maximal posterior probability is difficult, participants may fail to discover the best explanation for a particle motion even while expressing a clear preference for that explanation on evaluation trials.

1.3. Inverse reasoning as a process-level account

The computational-level account of object discovery predicts that the reasoner will discover the explanation with the maximum posterior probability. In practice, however, discovering the best explanation can be a computationally intractable problem. A consideration of processing constraints suggests that inverse reasoning might be more realistically formulated as a process-level rather than a computational-level account of object discovery. According to this formulation, the reasoner—being unable to consider every possible explanation of the data—considers a subset of the possible explanations and then uses inverse reasoning to evaluate the selected explanations (see the lower half of Fig. 2). The core prediction of this process-level inverse reasoning account is that people evaluate the proposed explanations by performing forward inferences; the task of explaining how people propose explanations is set aside as a problem to be solved by some other account. It is not clear what form this other account might take, but there are a few reasonable candidates. For example, one possibility is that people learn inferential rules that allow them to generate potential explanations from the given observations. Another possibility is that domain-general search algorithms allow people to identify hypotheses that have high posterior probabilities. Statisticians and computer scientists have identified various search algorithms that generate possible explanations in a manner that allows them to approximate computational-level inverse reasoning, and rational process models based on these search algorithms have been proposed as psychological accounts of backward inference (Griffiths, Vul, & Sanborn, 2012).

Unlike the computational-level inverse reasoning account, the process-level account allows for the possibility that people will fail to discover the best explanation for the observed data. In the context of our experiments, therefore, the process-level account allows for the possibility that a participant will fail to identify the best explanation on a discovery trial even while preferring that explanation on evaluation trials. Because the process-level account proposes that people perform forward inferences during object discovery, however, it still predicts that people will only provide explanations that actually support a forward inference to the observations. To test this prediction, we asked the participants in our experiments to predict the motion of a particle given their own explanations from the discovery trials. Both formulations of the inverse reasoning account propose that the predicted motions on these trials will closely resemble the to-be-explained motion from the corresponding discovery trials.

The process-level formulation of the inverse reasoning account has inspired computational models of human reasoning (e.g., Brown & Steyvers, 2009; Sanborn, Griffiths, & Navarro, 2010; Ullman, Goodman, & Tenenbaum, 2010), and is also supported by the observation that there is close agreement between forward and backward inferences on some inferential tasks (Fernbach, Darlow, & Sloman, 2010, 2011). Fernbach et al. (2010, 2011), for example, found close agreement between participants' estimates of (1) the probability that an infant with a drug addiction had a mother with a drug addiction (a backward inference), (2) the probabilities that mothers with and without drug addictions would give birth to infants with drug addictions (forward inferences), and (3) the prevalence of drug-addicted mothers (the prior plausibility of the focal hypothesis). This agreement is exactly what would be expected if people's backward inferences involved inverse reasoning.[2]

Other evidence, however, suggests that forward and backward inference may be supported by distinct psychological processes. Medical (e.g., Patel & Groen, 1986) and physics (e.g., Larkin, McDermott, Simon, & Simon, 1980) experts, for example, often apply inferential rules that directly map observed effects onto unobserved causes. These rules allow these experts to make backward inferences without carrying out any kind of forward inference. In the context of our object-discovery task, heuristic rules might similarly allow a reasoner to generate explanations directly from the observed particle motions. For example, one heuristic generates potential explanations by placing an attractor directly along the path of a particle's motion. This heuristic often produces reasonable explanations, but there are also situations where it produces explanations that do not actually predict the observed motion. In such situations, it might be possible to observe a dissociation between forward and backward inference in which people provide explanations that do not actually explain the particle motion. This dissociation would be inconsistent with both formulations of the inverse reasoning account.

1.4. Summary

Inverse reasoning can be formulated as either a computational- or process-level account of object discovery. Because the computational-level formulation specifies which explanations a person will produce and the process-level formulation specifies how these explanations are identified, these formulations are to some extent independent. Table 1 includes examples to which both, neither, or just one of these formulations applies. Rational process models are process-level formulations of the inverse reasoning account, and these implementations might successfully or unsuccessfully approximate computational-level inverse reasoning (see rows one and three in Table 1). Backward inference could also be implemented by applying inferential rules that directly map the observed effects to the unobserved causes. Such an implementation—which would not involve inverse reasoning at the process-level—might be either consistent or inconsistent with computational-level inverse reasoning (see rows two and four in Table 1). For example, experienced doctors might rely on inferential rules that allow them to correctly diagnose a disease from its symptoms, but novice doctors might apply inappropriate heuristics and make incorrect diagnoses.

We developed three experiments with the aim of evaluating both the computational-level and the process-level formulations of the inverse reasoning account. Experiments 1 and 2 tested the computational-level inverse reasoning account by comparing our participants' inferences on discovery and evaluation trials. Experiment 3 tested the process-level inverse reasoning account by comparing participants' inferences on discovery and prediction trials.

2. Experiment 1

Our primary goal in Experiment 1 was to test whether our participants would discover the best explanations of the presented particle motions. To do so, we first identified what we believed to be the best explanation of each of the particle motions from the discovery trials. The simplest possible observed motion is a straight line, and the obvious explanation for this trajectory is that the particle is either moving toward an attractor or moving away from a repeller. The particle motions presented in the discovery phase included some of the next simplest cases. The discovery phase contained three focal scenes that could be explained in at least two ways (see Figs. 1 and 3). First, each scene had a parsimonious explanation that invoked a relatively small number of stationary attractors and repellers (see the left column of Fig. 3). For example, the wall motion could be explained by positing a single repeller near the left wall of the arena: this repeller would explain the initial diagonal motion of the particle, and so long as one assumes that friction along the wall is minimal, it would also explain the subsequent horizontal motion of the particle. In our judgment, these parsimonious explanations were among the best explanations for the particle motions. Second, each scene had a non-parsimonious alternative explanation, where the attractors and repellers spontaneously appear, disappear, or move, or where the attractors and repellers have different strengths.

The computational-level inverse reasoning account predicts that the parsimonious explanations should be generated during the discovery phase and enthusiastically endorsed during the evaluation phase, and we were especially interested in testing this prediction. An alternative possibility is that some participants would discover explanations by applying a heuristic in which an attractor or repeller is always placed along the path of the particle. If

[2] Fernbach, Darlow & Sloman refer to inferences from causes to effects as *predictive* and inferences from effects to causes as *diagnostic*, and other researchers have used the same terminology (Waldmann & Holyoak, 1992).

Table 1
Examples of reasoning systems that are consistent or inconsistent with the computational-level and process-level formulations of the inverse reasoning account.

Consistency		Example
Computational	Process	
Yes	Yes	Rational process model that relies on a good approximation
Yes	No	Expert using "compiled" inferential rules
No	Yes	Rational process model that relies on a poor approximation
No	No	Novice using flawed heuristic rules

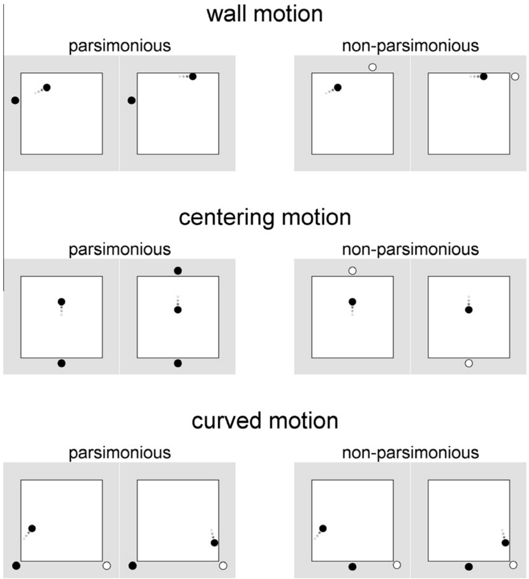

Fig. 3. Parsimonious and non-parsimonious explanations for the focal motions in Experiment 1. Each explanation is represented as a specification of the locations of the attractors (white circles outside of the arena) and repellers (black circles outside of the arenas) in two different snapshots from the particle motion. The alternative explanation for the curved motion is considered non-parsimonious because it only explains the particle motion under the assumption that the attractor is stronger than the repeller; otherwise, the repeller would pin the particle to the leftmost wall of the arena.

some of our participants used this strategy, then they would fail to discover the parsimonious explanations even while expressing a preference for the parsimonious explanations in the evaluation phase—a result that would be inconsistent with the computational-level inverse reasoning account.

Experiment 1 also included a prediction phase in which participants predicted the particle motions that would result from several kinds of configurations of the attractors and repellers. Each presented configuration could be viewed as an explanation for one of the motions observed during the discovery phase. Some of the prediction trials presented participants with the parsimonious explanations for the three focal particle motions. These trials allowed us to determine 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. On other prediction trials, participants predicted the expected path of the particle given their own explanations for the particle motions from the discovery phase. This allowed us to assess whether the explanations generated in the discovery phase would in fact explain the observed particle motion according to the participants' own forward inferences.

2.1. Method

2.1.1. Participants

Thirty undergraduates at Carnegie Mellon University participated for course credit.

2.1.2. Materials and procedure

Participants were asked to imagine themselves working for a scientist who studies the motion of "particles" within a rectangular arena. Participants learned that the particle motions were caused by "attractors" and "repellers" which were always located outside the arena. Participants then viewed three scenes that were designed to illustrate the basic properties of the attractors and repellers. All scenes were displayed as a sequence of bird's-eye-view camera snapshots of the particle motion. The first two snapshots of each scene always showed the particle being restrained by a "holder" box that prevented the particle from moving; these initial snapshots established that the particle did not have an initial velocity.

Before beginning the experiment, participants viewed one familiarization scene showing that particles move toward attractors and another familiarization scene showing that particles move away from repellers. The attractors and repellers were depicted as green and red circular objects, respectively. A third familiarization scene showed that a particle placed between two attractors moved toward the closer one, and the accompanying instructions explained that nearby attractors and repellers exert greater forces on the particle than more distant attractors and repellers. After completing the familiarization trials, the participants completed (in order) the discovery, prediction, and evaluation phases.

2.1.2.1. Discovery phase. In the discovery phase, participants were told that the camera was not set up properly during some of the experiments. As a result, the snapshot sequences from these experiments failed to capture the locations of the attractors and repellers. Participants were asked to infer the locations of the unseen attractors and repellers by reviewing the snapshot sequences.

Participants completed a practice trial and then generated explanations for fifteen scenes in which the attractors and repellers were not visible. We will focus on the three critical scenes that are depicted in Fig. 1. In the wall-motion scene, 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 centering-motion scene, the particle moved from the center of the arena to the top wall, paused, and then returned to the center of the arena. In the curved-motion scene, the particle moved along a curved path from the lower-left corner of the arena to the lower-right corner of the arena. Because the remaining twelve scenes did not have analogues in the prediction or evaluation phases, we do not discuss them further.[3]

On each discovery trial, the participants viewed the sequence of snapshots showing the particle motion. Then participants were asked to explain the particle motion by specifying where the attractors and repellers would have been in two of the snapshots (see the outlined snapshots in Fig. 1). The instructions explained that the participants were reporting the locations of the attractors and repellers in two separate snapshots 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 snapshots and a summary of the to-be-explained particle motion. Participants could place attractors and repellers by clicking on any location outside the arena, and they were able to move and erase the attractors and repellers as needed. A "reuse" button located between the two response pictures copied the attractors and repellers in the first response snapshot to the corresponding locations in the second response snapshot. Participants were encouraged to provide up to three explanations for each scene; each explanation was entered on a separate screen. Participants were also encouraged to supplement their pictorial explanations with written explanations as needed. These supplementary explanations were entered in a text box that appeared when the participant clicked the "add explanation" button. Because supplementary explanations were uncommon, they are not discussed further.

After providing explanations for a particle motion, 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 or explanations, but these ratings were not further analyzed. The ratings of the provided explanations were used to identify each participant's *preferred explanation*, which we defined as the explanation that received the highest rating, with ties broken by selecting the explanation that was provided first. These preferred explanations were presented to the participants in the prediction and evaluation phases.

2.1.2.2. Prediction phase. Participants were asked to predict the particle paths given the locations of the attractors and repellers. Some of the prediction trials presented participants with configurations corresponding to an early snapshot from each of the parsimonious and non-parsimonious explanations (see the first snapshot for each explanation in Fig. 3). For explanations where the configuration of the attractors and repellers changed during the particle motion, participants also predicted the motion of the particle given the configuration of the particles in a later snapshot (see the second snapshots for each relevant explanation in Fig. 3).

Three other prediction trials presented each participant with configurations of the attractors and repellers taken from his or her own explanations for the three focal discovery scenes. In particular, for each focal discovery scene, we presented each participant with the first response snapshot from his or her preferred explanation. These trials allowed us to assess whether the explanations that the participants provided in the discovery phase would have produced the observed motion according to the participants' own forward inferences. Four additional prediction trials served as pilot trials for future experiments and will not be discussed further.

2.1.2.3. Evaluation phase. In the evaluation phase, participants once again viewed the wall, centering, and curved motions. There were two evaluation trials for each motion scene: on one trial, the participant chose between his or her preferred explanation and the parsimonious explanation shown in the left column of Fig. 3; on the other trial, the participant chose between his or her own explanation and the non-parsimonious explanation (see the right column of Fig. 3). The non-parsimonious explanations for the wall and centering motions involved more configuration changes than the corresponding parsimonious explanations and the alternative explanation for the curved motion only explained the particle motion if the attractor is assumed to be stronger than the repeller. The trials with the non-parsimonious explanations served to control for any task demands associated with asking a participant to choose between his or her explanation and an experiment-provided explanation. To further limit task demands, the explanations were described as responses provided by other participants.

On each evaluation trial, participants were asked to choose the explanation that was more likely to be the true explanation. Because participants occasionally generated the parsimonious or non-parsimonious explanations themselves, this meant that participants were sometimes presented with a choice between two identical explanations. For these situations, participants were provided with a "these explanations are identical" button.

2.2. Results

The main finding is that participants often failed to generate parsimonious explanations in the discovery phase but frequently endorsed them in the evaluation phase. This result suggests that inferences in the discovery phase were incompatible with the computational-level inverse reasoning account. Before reviewing the evidence for this finding in full detail, we first provide a general overview of the explanations that participants generated in the discovery phase.

2.2.1. Overview of discovery responses

Fig. 4 depicts the most common preferred explanations for the focal discovery trials. "Heuristic" explanations that posited attractors or repellers directly on the path of the

[3] Most of these scenes were slight variations of the focal scenes for which the parsimonious explanations were inappropriate. We included these scenes to serve as fillers and to show that any participants who generated the parsimonious explanation understood its scope and the circumstances under which it applied. As we discuss in the Results section, however, participants rarely generated the parsimonious explanation. As a consequence, these remaining scenes were generally uninformative.

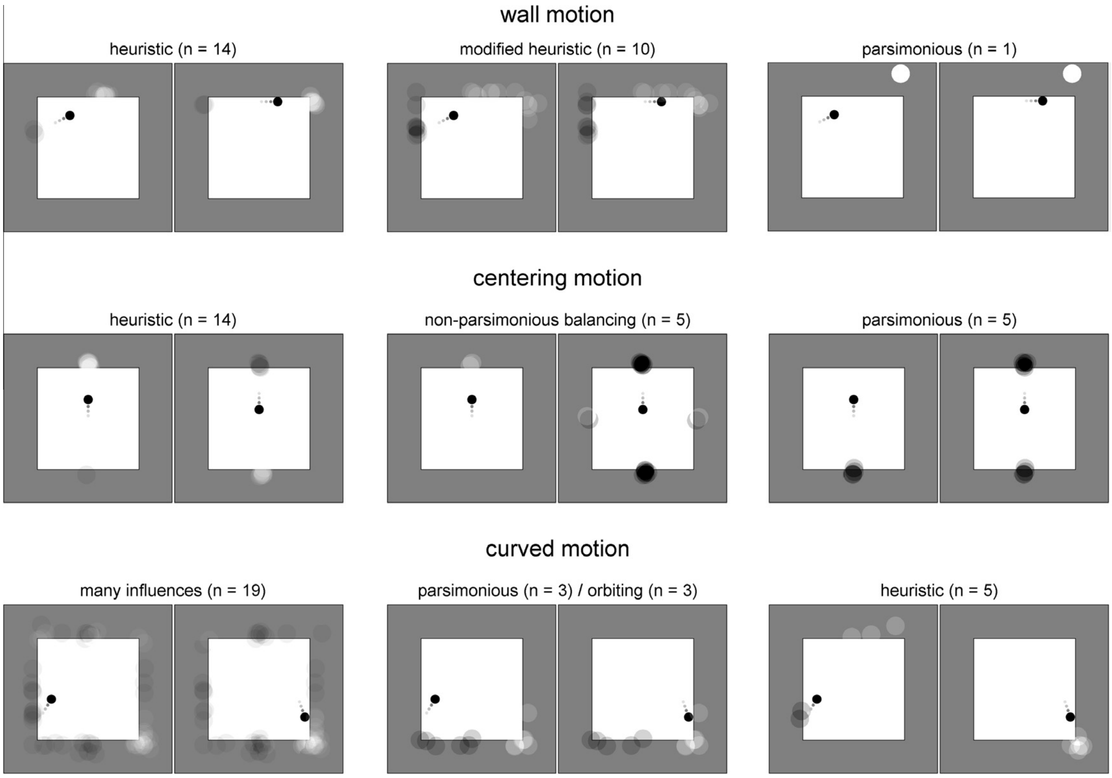

Fig. 4. "Heatmaps" that represent the most common preferred explanations for each discovery scene. Locations where attractors were often placed are depicted as brighter areas and locations where repellers were often placed are depicted as darker areas. The number of responses represented by each explanation type is shown in parentheses.

particle were common explanations for the wall and centering motions; heuristic explanations were less common but were still represented among the explanations for the curved motion. For the wall motion, many of the other explanations resembled the heuristic explanation but posited additional attractors or repellers (the "modified heuristic" responses). Often, these modified explanations posited that the attractors and repellers along the paths of the particle motion remained stationary throughout the particle motion. Only a single participant provided a parsimonious explanation. For the centering motion, many of the non-heuristic explanations posited balancing repellers in the second response snapshot. Some of these explanations were parsimonious (the "parsimonious" explanations), but others posited an additional attractor in the first response picture (the "non-parsimonious balancing" explanations). For the curved motion, most participants provided explanations that posited the simultaneous presence of many attractors and repellers (the "many influences" explanations). Other participants provided explanations that posited a single stationary repeller and a single stationary attractor. These explanations were classified as "parsimonious" when the repeller was located at the left of the arena and "orbiting" when the repeller was located at the bottom of the arena. (As explained previously, the orbiting explanation only explains the particle motion under the assumption that the attractor is stronger than the repeller. Therefore, it was not considered parsimonious.)

2.2.2. Inverse reasoning as an account of hypothesis discovery

To investigate the generation of parsimonious explanations, we classified each explanation as parsimonious or non-parsimonious. An explanation of the wall motion was coded as parsimonious when it posited a single stationary attractor or repeller in a location that would explain the particle motion. An explanation of the centering motion was coded as parsimonious when it invoked a single repeller in the first response snapshot and balancing repellers in the second. An explanation of the curved motion was coded as parsimonious when it posited a static configuration of two hidden objects that would explain the particle motion without any additional assumptions (e.g., without assumptions about the relative strength of an attractor and repeller). Using these criteria, we found that only 2, 6, and 3 participants generated a parsimonious explanation for the wall, centering, and curved motions, respectively.[4]

This finding is problematic for the computational-level inverse reasoning account, which predicts that participants should have found and provided the best explanations for

[4] The number of participants who were classified as providing parsimonious explanations exceeds the number of parsimonious explanations that are depicted in Fig. 4. The reason is that some participants provided the parsimonious explanations as secondary rather than preferred explanations (Fig. 4 only depicts preferred explanations).

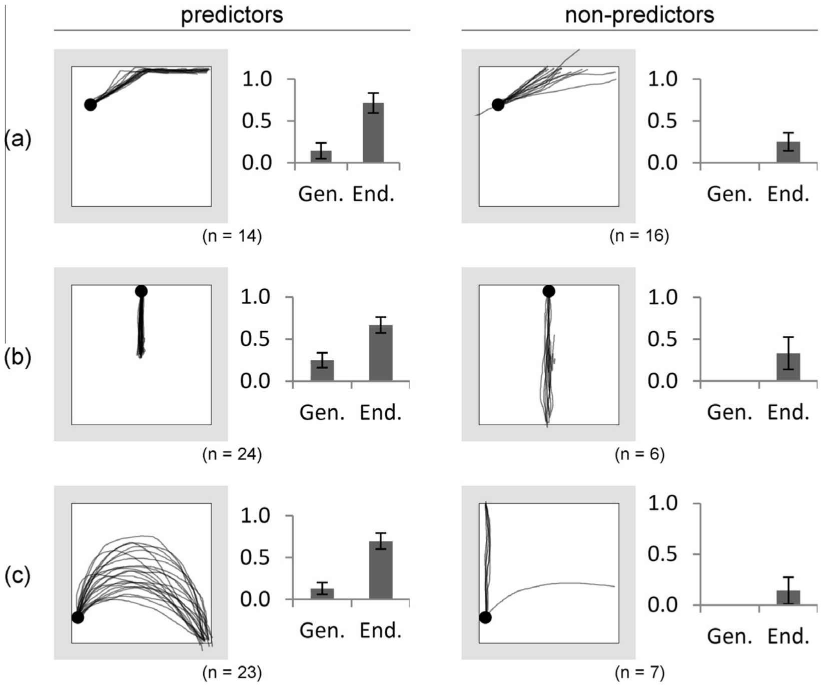

Fig. 5. Inferences about the parsimonious explanations for the (a) wall, (b) centering, and (c) curved motions. The path traces within the arenas show the predicted particle motions for the parsimonious explanations; note that participants were classified as "predictors" and "non-predictors" depending on whether their predicted particle motions resembled the actual particle motions from the discovery phase (all classifications were made by the first author). The bar graphs show the rates at which participants generated and endorsed the parsimonious explanations in the discovery and evaluation phases. Gen. = generated; End. = endorsed.

the particle motions. The inverse reasoning view might be reconciled with this finding if we suppose that participants did not actually believe that the parsimonious explanations were the best explanations. The responses of the participants on the prediction and evaluation phases undermine this supposition, however. Fig. 5 shows that many participants agreed that the parsimonious explanation would indeed produce the observed motion: when provided with the parsimonious explanations, 14, 24, and 23 out of 30 participants were "predictors" who predicted motion paths that closely resembled the wall, centering, and curved motions, respectively. Note also that even among the predictors, participants rarely generated the parsimonious explanations.

The evaluation phase further confirmed that many participants recognized the merits of the parsimonious explanations. Recall that some of the trials in the evaluation phase asked participants to choose between a parsimonious explanation and their self-generated explanations. Participants endorsed parsimonious explanations on these trials at much greater rates than they generated them in the discovery phase: 14 out of 30, 18 out of 30, and 17 out of 30 participants endorsed the parsimonious explanation for the wall, centering, and curved motions, as compared to the 2 out of 30, 6 out of 30, and 3 out of 30 participants who generated parsimonious explanations in the discovery phase. In addition, the rate of endorsement was even higher when the analysis is restricted to the participants who believed that the parsimonious explanations were valid (see Fig. 5).[5] Among the predictors, participants were marginally more likely to endorse a parsimonious explanation for the wall motion than to generate a parsimonious explanation for the wall motion ($p = .054$) and significantly more likely to endorse than to generate a parsimonious explanation for the centering ($p = .008$) and curved ($p < .001$) motions (all by Fisher's exact test).

It is possible that the non-predictors in Fig. 5 endorsed explanations that they did not generate because they relied on different physical theories during the discovery and evaluation phases. For example, 13 out of 16 wall-motion non-predictors predicted that the particle would stop moving after it collided with the upper wall, and this prediction presumably reflects an assumption that the force of friction between the particle and the wall is significant. During the evaluation phase, however, participants were asked to consider a parsimonious explanation that suggests that the friction produced by contact with the wall is not significant. Some of the non-predictors may have endorsed this explanation after inferring that their initial assumptions about friction were incorrect. Non-predictors

[5] For this analysis, participants were counted as having endorsed a parsimonious explanation when either (a) the participant preferred the experimenter-provided parsimonious explanation to his or her own or (b) the participants' own explanation was parsimonious. In the latter case, the evaluation trial involved a choice between two parsimonious explanations, so any response constituted the endorsement of a parsimonious explanation.

who revised their physical theories in this way would generate incompatible responses to the discovery and evaluation phases even if they relied on inverse reasoning during both phases.

Critically, however, the responses of the predictors in Fig. 5 cannot be explained in this way. Even if these predictors were relying on physical theories different from the one that we had in mind, their predictions suggest that the parsimonious explanations were valid with respect to these physical theories. The fact that many of these predictors endorsed but did not generate the parsimonious explanations suggests that the computational-level version of the inverse reasoning account is limited as an account of hypothesis discovery.

One remaining concern is that the high endorsement rates from the evaluation phase may reflect a task demand implicit in asking the participants to choose between their own explanations and experimenter-provided explanations. Participants, however, often preferred their own explanations to competing explanations in other situations. This can be seen, for example, in the low rates at which the non-predictors endorsed the parsimonious explanations (see Fig. 5). It can also be seen by comparing evaluation trials that involved parsimonious competing explanations to evaluation trials that involved non-parsimonious explanations from Fig. 3. Participants were more likely to prefer parsimonious explanations to their own than to prefer non-parsimonious explanations to their own on the evaluation trials for the wall (13 of 30 vs. 11 of 30 participants), centering (13 out of 30 vs. 3 out 30 participants), and curved (15 out of 30 vs. 7 out of 30 participants) motions. These differences were significant after collapsing across the three discovery scenes, $p = .003$ by Fisher's exact test.

2.2.3. Forward inferences from preferred explanations

Both formulations of the inverse reasoning account posit that a good explanation should support a strong forward inference to the observed particle motion. Recall that three prediction trials tested this prediction by presenting participants with their preferred explanations for the particle motions. The inverse reasoning account predicts that the predicted motions on these trials should closely resemble the actual motions from the discovery phase, at least during the initial stages of the particle's motion (the predicted and actual motions might reasonably diverge in later stages of the motion if the participant's explanation involved a changing configuration of attractors and repellers). Fig. 6 shows that even though the predicted and actual motions were often consistent, these motions were inconsistent in 17 out of 90 cases. There are two possible interpretations of this finding. First, people may have provided "explanations" that are simply invalid, even according to their own forward inferences. If so, then this would provide evidence for a fundamental dissociation between forward and backward inference. The second interpretation is more mundane: perhaps the discrepancies between the discovery and prediction trials are explained by differing assumptions about the initial velocity of the particle. In particular, given that the particle had already moved a short distance before the first response snapshot in the discovery phase, it was natural to assume that the particle was already in motion at that point. In contrast, there was no reason to assume that the particle was in motion at the outset of the prediction phases. In summary, although the experimental results are consistent with the existence of a fundamental dissociation between forward and backward inference, more mundane interpretations are also possible. We sought to find more conclusive evidence for this dissociation in Experiment 3.

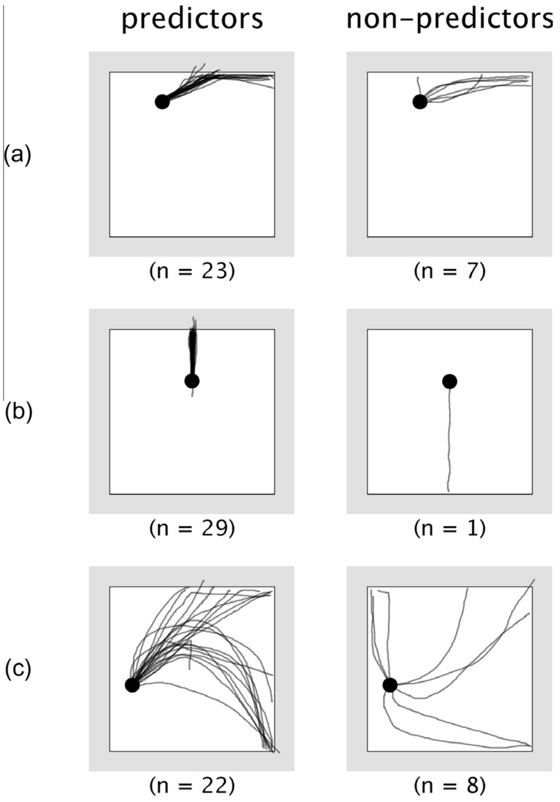

Fig. 6. The predictions of the participants when provided with their own explanations for the (a) wall, (b) centering, and (c) curved motions. Participants were classified as "predictors" when they predicted particle motions that initially resembled the motions from the discovery trials.

2.2.4. Inverse reasoning as an account of hypothesis evaluation

The inverse reasoning account fares better as an account of hypothesis evaluation than as an account of hypothesis discovery. The inverse reasoning account predicts that participants should prefer explanations with higher posterior probabilities on the evaluation trials, and participants' inferences were almost always consistent with this prediction. This is revealed in part by the observation that inferences on the prediction and evaluation trials were—as the inverse reasoning account predicts—closely related: the predictors very often endorsed the parsimonious explanation and the non-predictors almost never did (see the endorsement rates in Fig. 5). These differences were significant after collapsing across the discovery trials, $p < .001$ by Fisher's exact test. Furthermore, as noted previously, participants were more likely to prefer a parsimonious explanation to their own explanation than to prefer one of the less parsimonious alternative explanations to their own explanation.

2.3. Discussion

Our main experimental finding is that even though participants frequently endorsed the parsimonious explanations in the evaluation phase, they rarely provided them in the discovery phase. This suggests that our participants failed to consider the parsimonious explanations in the discovery phase. In some respects, this is not surprising. Given that there are an infinite number of possible explanations for any given particle motion, it is impossible for participants to consider every possible explanation. Yet the parsimonious explanations were not especially complicated or obscure; rather, they were simple explanations that should have been within reach. So why did participants fail to generate them? Experiment 1 was not designed to address this question, but one possibility is that our participants "satisficed" by prematurely terminated the search for better explanations after finding the heuristic explanations. Of course, it is also possible that participants were not able to discover the parsimonious explanations despite searching for them: perhaps the parsimonious explanations were not as simple as we expected. To decide between these possibilities, we conducted Experiment 2, which was designed to explore whether the parsimonious explanations were within the reach of our participants.

3. Experiment 2

Experiment 2 was designed to demonstrate that people are able to discover the parsimonious explanations for the discovery scenes when the heuristic explanations for those scenes are unavailable. To manipulate whether the heuristic explanations were available to participants in Experiment 2, we placed additional constraints on what sorts of explanations were allowed. In the *static* condition, participants were instructed that attractors and repellers were always stationary and that they could neither appear nor disappear. Participants in the static condition were also informed that there were limits on the number of attractors and repellers that could be involved in the explanation of each scene. These constraints were designed to exclude the heuristic explanations for the wall, centering, and curved motions that were presented in Experiment 2. In the wall motion, for example, heuristic explanations posit that one attractor or repeller is present during the particle's diagonal motion and that a different attractor or repeller is present during the particle's motion along the wall. This explanation was not available to participants in the static condition because it involved disappearing and appearing attractors and repellers. The *dynamic* condition served as a control condition in which the heuristic explanations were available. Participants in the dynamic condition could provide explanations that posited any number of appearing, disappearing, and moving attractors and repellers. We expected that many participants in the static condition would provide the parsimonious explanations for the particle motions. Such a finding would indicate that the parsimonious explanations are within the reach of our participants and that participants in Experiment 1 could have discovered them had they only looked for them.

3.1. Method

Except where noted, our method was identical to the method from Experiment 1.

3.1.1. Participants

Thirty-eight undergraduates at Carnegie Mellon University participated for course credit and were randomly assigned to the static ($n = 19$) and dynamic ($n = 19$) conditions.

3.1.2. Materials and procedure

The three particle motions of interest are shown in Fig. 7. The wall and curved particle motions represent minor variations on the corresponding particle motions from Experiment 1.[6] The differences between centering motions in the two experiments were more significant: in Experiment 1, the particle traveled to the top of the arena and then to the center of the arena; in Experiment 2, the particle traveled from the top of the arena to the center of the arena (i.e., the motion involved only the latter half of the motion from Experiment 1). The reason for this alteration was to allow the particle motion to be explained without positing a changing configuration of the attractors and repellers.

In addition to the three particle motions depicted in Fig. 7, there were six filler particle motions. Because these filler trials are not relevant to our primary experimental question, we do not discuss them further.

3.1.2.1. Discovery phase. After viewing a particle motion using the same interface as in Experiment 1, participants were asked to provide the single "most likely" explanation for the particle motion. The response interface was modified so that participants in the dynamic condition were able to specify the exact configuration of attractors and repellers in every snapshot. The response interface displayed a single snapshot at a time, and participants were able to navigate to different snapshots using forward and backward buttons. Participants were able to place and erase attractors and repellers in any snapshot, but the effect of these actions differed between the experimental conditions. When a participant in the dynamic condition placed an object in a snapshot, the object was automatically placed in all subsequent snapshots. Likewise, when a participant in the dynamic condition erased an object, it was automatically removed from

[6] Some of the changes were intended to make the parsimonious explanations more appealing. By decreasing the angle of the collision between the wall and the particle in the wall-motion scene, for example, we intended to limit the influence of friction forces between the particle and the wall. Other changes were intended to render some of the non-parsimonious explanations less appealing. For example, reducing the height of the arena in the curved-motion scene meant that a repeller placed along the bottom wall of the arena would not be at the focal point of the curve formed by the particle's path.

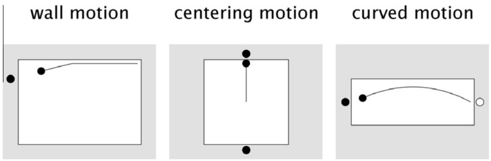

Fig. 7. The particle motions and parsimonious explanations for the wall, centering, and curved motions. The circles inside the arenas represent the initial locations of the particles, and the path traces show where the particle moved during the scene. The circles outside of the arenas represent the locations of the repellers (black circles) and attractors (white circles) for the parsimonious explanations that were presented in the prediction and evaluation phases.

all subsequent snapshots. The interface therefore allowed participants in the dynamic condition to specify when an attractor or repeller appeared and disappeared. By placing an attractor in the third snapshot and erasing it in the tenth snapshot, for example, a participant in the dynamic condition could specify that the attractor was present in snapshots three through nine. When a participant in the static condition placed or erased an object, it was placed or erased from all snapshots. As a consequence, participants in the static condition were only able to provide explanations that posited a static configuration of attractors and repellers.

The response interface also prevented participants in the static condition from providing explanations that contained more than a certain number of hidden objects. Explanations for the centering and curved motions could posit up to two hidden objects and explanations for the wall motion could posit one hidden object. The instructions emphasized that the allowable number of hidden objects represented an upper limit rather than a target: participants were encouraged to provide the best possible explanation for the particle motion, regardless of whether or not it contained fewer than the allowable number of hidden objects. For participants in the static condition, the response interface displayed the number of hidden objects that had been placed and the number of additional hidden objects that were available for placement.

To ensure that participants understood how placing or erasing a hidden object in one snapshot influenced whether the hidden object was present in other snapshots, participants were shown summaries of their explanations as they entered them. The summaries showed the posited locations of the attractors and repellers across a sequence of snapshots that spanned the particle motion. Additionally, participants were shown another summary of the explanation after submitting it. Upon viewing this summary, participants were asked to confirm that the provided explanation was the explanation that they had intended to submit. Participants who wanted to make changes to their explanations during the confirmation trial were allowed to return to the response interface to do so.

In contrast to Experiment 1, participants were not asked to rate the provided explanations. Furthermore, participants were also provided with a "no explanation" button that allowed them to indicate when they were unable to find an explanation for the particle motion.

3.1.2.2. Prediction phase. In the prediction phase, participants predicted the motion of the particle given various configurations of the attractors and repellers. Three of the configurations corresponded to the parsimonious explanations shown in Fig. 7.

3.1.2.3. Evaluation phase. The trials in the evaluation phase required participants to choose between the parsimonious explanations for the focal scenes and their own explanations for those scenes. Some participants were unable to generate explanations for some of the motions in the discovery phase, and those participants were not asked to choose between explanations for those motions.

3.2. Results

Our primary finding is that participants in the static condition often discovered parsimonious explanations for the discovery scenes but that participants in the dynamic condition rarely did so. This result suggests that the parsimonious explanations were in fact accessible and that participants in the dynamic condition could have discovered them had they searched beyond the few hypotheses that came immediately to mind.

3.2.1. Discovery phase

Fig. 8 depicts the most common explanations for each discovery scene and shows the number of participants in each condition who generated each explanation. Before discussing the responses in detail, we first note that participants in the static condition discovered parsimonious explanations on 31 out of 57 discovery trials but that participants in the dynamic condition discovered parsimonious explanations on only 12 out of 37 discovery trials. This difference was statistically significant, $p < .001$ by Fisher's exact test, and consistent across each of the discovery scenes, p's = .017, .099, and .090 for the wall, centering, and curved motions, respectively (all by Fisher's exact test).

As shown in Fig. 8, the most common explanations for the wall motion were "parsimonious", "heuristic", and "simple-but-invalid" explanations. The parsimonious explanation posited a single hidden object in a location that would explain the entire particle motion, the heuristic explanations posited a changing configuration of the attractors and repellers, and the simple-but-invalid explanations posited a single hidden object in a location that did not explain the particle motion. The remaining participants provided less common explanations ($n = 0$ in the static condition and $n = 3$ in the dynamic condition) or indicated that they could not find an explanation for the particle motion ($n = 2$ and $n = 0$ in the static and dynamic conditions).

The prevalence of the simple-but-invalid explanations raises the possibility that some participants preferred providing an incomplete explanation to providing no explanation at all. Other considerations suggest, however, that this preference—to the extent that it existed—was not widespread. The experimental instructions encouraged participants to use the "no explanation" response whenever appropriate, and many participants used this option at

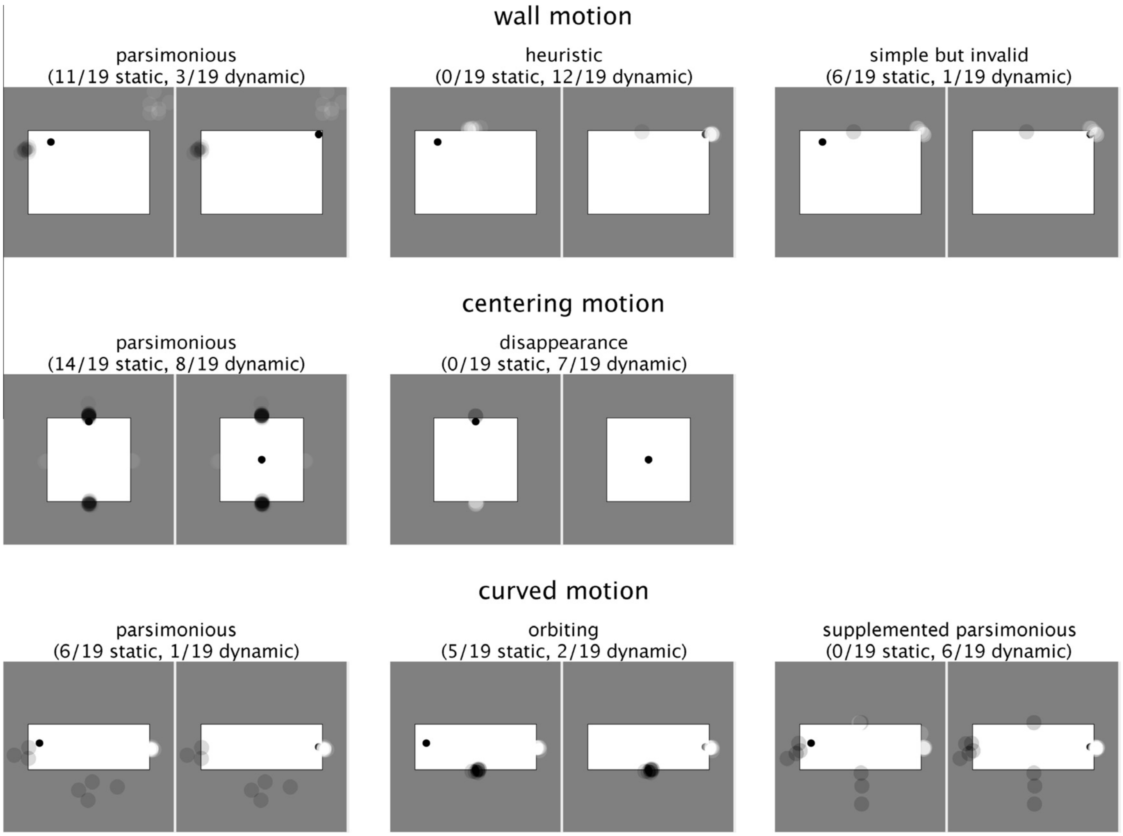

Fig. 8. These heatmaps show the most common explanations for the wall, centering, and curved motions. For each explanation, the first and second heatmaps correspond to the initial and final configurations of the attractors and repellers. The number of responses represented by each explanation type is shown in parentheses.

some point in the experiment: 14 out of 19 and 5 out of 19 participants did so in the static and dynamic conditions, respectively.

The most common explanations for the centering motion were "parsimonious" explanations, which posited two repellers at the top and bottom of the arena or two attractors at the left and right of the arena, and "disappearance" explanations, which posited a single attractor or repeller that disappeared as the particle approached the center of the arena. The remaining participants either provided uncommon explanations ($n = 3$ and $n = 4$ in the static and dynamic conditions) or failed to find an explanation ($n = 2$ and $n = 0$ in the static and dynamic conditions).

The most common explanations for the curved motion were "parsimonious" explanations. Some of these explanations posited a repeller along the left wall of the arena, and the rest posited a repeller below the bottom wall of the arena. Critically, the repeller in these remaining explanations was displaced some distance from the bottom wall of the arena. Because this displacement would have led the repeller to exert less of its force in the horizontal plane, we classified these explanations as parsimonious even while classifying the standard "orbiting" explanations as non-parsimonious. (Recall that the standard "orbiting" explanation is only valid under the assumption that the attractor is stronger than the repeller.) The "supplemented parsimonious" explanations resembled the parsimonious explanations but posited additional attractors and repellers. The remaining participants either provided uncommon explanations ($n = 5$ and $n = 8$ in the static and dynamic conditions) or failed to find an explanation ($n = 3$ and $n = 2$ in the static and dynamic conditions).

3.2.2. Prediction phase

Fig. 9 shows the predictions of participants given the parsimonious explanations of the wall, centering, and curved motions. As was the case in Experiment 1, participants usually regarded the parsimonious explanations as valid: most of the predicted motions closely resembled the observed motions from the discovery phase. As expected, the proportion of participants who were classified as "parsimonious predictors" was similar across the static and dynamic conditions for any of the prediction trials, all p's > .50 by Fisher's exact test. The predictions of the non-predictors usually corresponded to the predictions of the non-predictors in Experiment 1: although the non-predictors did not expect the parsimonious explanations to produce the particle motion from the discoverable scenes, their predictions often reflected reasonable alternative assumptions about the forces involved in the motions.

The prediction trials also address the concern that participants in the static condition provided the parsimonious explanation only because they were prevented from providing other explanations. The bar graphs in Fig. 9 show that even when the analysis is restricted to participants who viewed the parsimonious explanations as valid (i.e., the predictors), participants in the static condition remained much more likely to discover (or "generate") a parsimonious explanation than participants in the dynamic condition, $p < .001$ by Fisher's exact test.

A final noteworthy finding is that a clear majority of predictors discovered the parsimonious explanation in the discovery phase. Indeed, the predictors of the wall and centering motions almost always found a parsimonious explanation for the observed motion. This finding provides even stronger evidence that the parsimonious explanation could have been generated by most of the participants in the dynamic condition and by most of the participants in Experiment 1.

3.2.3. Evaluation phase

As expected, the parsimonious explanations were usually endorsed by the predictors and often rejected by the non-predictors (see the "endorsed" bars in Fig. 9). When choosing between explanations, therefore, our participants reasoned in a manner that was consistent with inverse reasoning.

3.3. Discussion

When asked to explain a particle motion, our participants often considered a surprisingly limited set of possible explanations. The primary evidence for this conclusion is that participants in the dynamic condition often failed to consider the parsimonious explanations—explanations that the participants easily could have discovered had they only searched for them. This finding presents a serious problem for the computational-level inverse reasoning account, which does not allow for the possibility that participants will fail to identify the best explanations for a particle motion.

The process-level inverse reasoning account, however, does allow for the possibility that people will fail to consider some explanations on discovery trials (see Fig. 2). Under this process-level formulation, inverse reasoning explains how people evaluate explanations even if it does not explain how people decide which explanations to consider in the first place. Yet there is reason to question even this narrow formulation. In particular, recall that participants in Experiment 1 occasionally generated explanations that seemed invalid according to their own forward inferences, which would be incompatible with both the computational- and process-level formulations of the inverse reasoning account. Experiment 3 sought to investigate this issue in greater detail.

4. Experiment 3

In Experiment 3, we investigated whether people can be induced to provide explanations that do not actually explain the particle motion. We hypothesized that this sort of situation might arise when people reflexively accept heuristically-generated explanations and fail to check the validity of those explanations by performing a forward inference. Some of the participants in Experiment 1, for example, explained the wall motion by positing an explanation similar to the "lure" explanation in the first row and third column of Fig. 10. Although this explanation may seem compelling at first glance, deeper reflection reveals that it is invalid: the simultaneous presence of the attractor and repeller would produce a curved motion rather than the linear one that was observed in the wall motion. Given the methodological limitations of Experiment 1, it was not possible to confirm whether the participants would have recognized this explanation as invalid when asked to make a forward inference. Our main objective in Experiment 3 was to show that people sometimes provide invalid explanations and only recognize the problems with these explanations when explicitly asked to carry out forward inferences. This finding would provide evidence against both the computational- and process-level formulations of the inverse reasoning account.

The first column of Fig. 10 shows the five focal scenes from the discovery phase in Experiment 3. The wall, centering, and curved motions were similar to the motions from the previous experiments. The centering motion in the present experiment, however, involved two particles moving toward the center of the arena. In addition, the scene for the centering motion was cut short: participants were informed that the camera malfunctioned after a certain snapshot and that no further snapshots were taken. The other two motions were novel. The split motion involved two particles that traveled in opposite directions, and the lane motion involved a particle that traveled upward and then leftward in two narrow lanes. The second column of Fig. 10 shows some of the simplest possible configurations that explain these particle motions. These explanations are analogous to the parsimonious explanations in the previous experiments, and we expected that participants would often fail to generate them in the discovery phase.

The discovery scenes were designed so that each scene had a simple but flawed "lure" response (see the third column of Fig. 10). Each lure corresponded to an invalid application of an otherwise reasonable heuristic. Note, for example, that the lures for the wall, centering, split, and lane motions (rows (a), (b), (d), and (e)) could have been generated by applying a heuristic in which a stationary hidden object is placed along each linear particle trajectory. Yet none of these lures would actually explain the discovery scenes: the wall- and centering-motion lures (rows "a" and "b") would have produced curved particle motions, and the split- and lane-motion lures would have produced particle motions in which the particles would have moved toward the closest attractor. The curved-motion lure (row "c") is only valid under the assumption that the attractor exerts a greater force on the particle than the repeller. Because participants in Experiment 3 were explicitly instructed that closer attractors and repellers always exerted a greater force on a particle than more

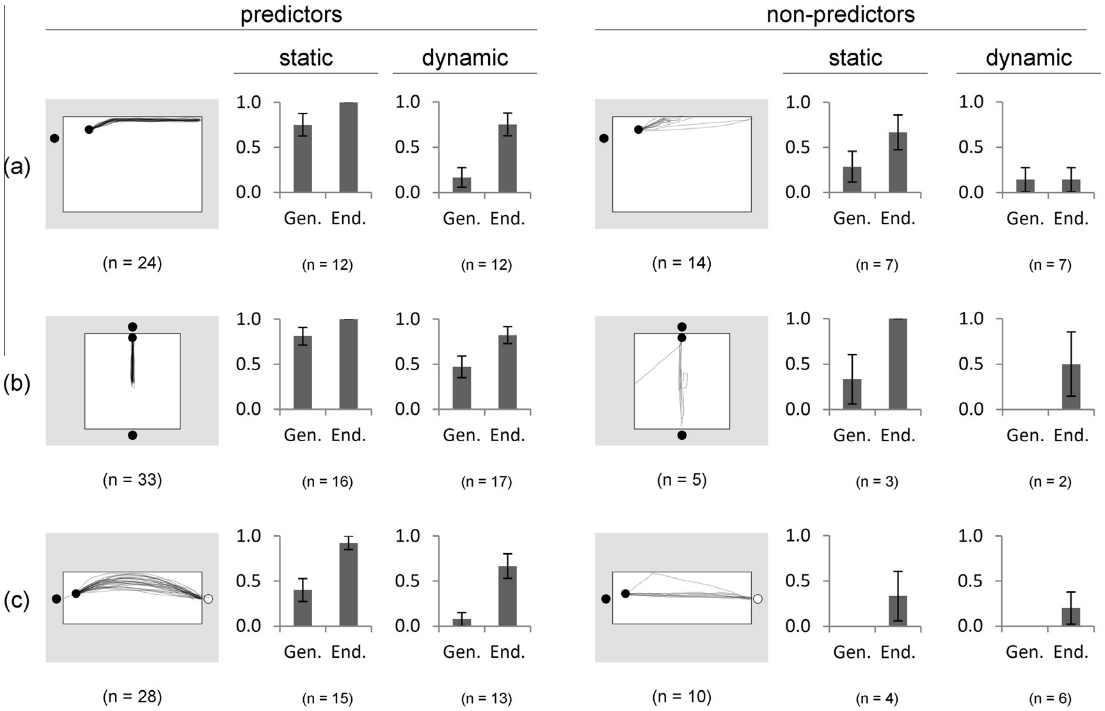

Fig. 9. Summary of the prediction and evaluation responses for the (a) wall, (b) centering, and (c) curved motions. For each scene, participants were classified as "predictors" or "non-predictors" depending on whether they predicted that the parsimonious explanation would produce the motion observed in the discovery phase. The arena figures show the predicted particle paths provided by the participants and the bar graphs show the proportion of participants who generated and endorsed parsimonious explanations in the discovery phase and evaluation phases, respectively. Gen. = generated; End. = endorsed.

distant attractors and repellers, this lure response was not a valid explanation in Experiment 3.

We expected that some participants would provide erroneous lure responses for the discovery scenes in Experiment 3. To confirm that these responses were invalid even according to the participants' own forward inferences, we asked participants to predict the motion of the particle given their own explanations for each discovery scene.

4.1. Method

Except where noted, the method was identical to the method from Experiment 2.

4.1.1. Participants

Twenty-three undergraduates from Carnegie Mellon University participated for course credit.

4.1.2. Materials and procedure

In addition to the five focal particle motions scenes depicted in Fig. 10, there were ten "filler" particle motions. To make it more difficult for participants in the prediction phase to simply recall the motion from the corresponding discovery scene, five of the filler motions were matched to the five focal motions: the particle motions in these matched scenes differed from those in the focal scenes, but both scenes began with the same initial particle configuration. We hoped that this design would force participants to make a genuine forward inference in the prediction phase.

4.1.2.1. Discovery phase. Participants viewed the five discovery scenes from Fig. 10 and ten filler discovery scenes. Participants were instructed that the configuration of attractors and repellers remained the same throughout every particle motion, and the response interface was modified accordingly. More specifically, rather than displaying specific snapshots from the relevant discovery motion, as in the previous experiments, the response display simply summarized the motion by showing the path that each particle followed, as well as the initial and final locations of the particles. Participants were asked to provide the single best explanation for the particle motion.

Note that the lure responses for the centering and split motions (rows "c" and "d" in Fig. 10) might be valid under the assumption that one or the attractors was "stronger" than the other. To exclude this possibility, participants were instructed that each hidden object had the same power and that a closer hidden object will always have a stronger influence on a particle than a more distant hidden object.

4.1.2.2. Prediction phase. In the prediction phase, participants made two predictions for each discovery scene. One of these predictions was made for a scene that

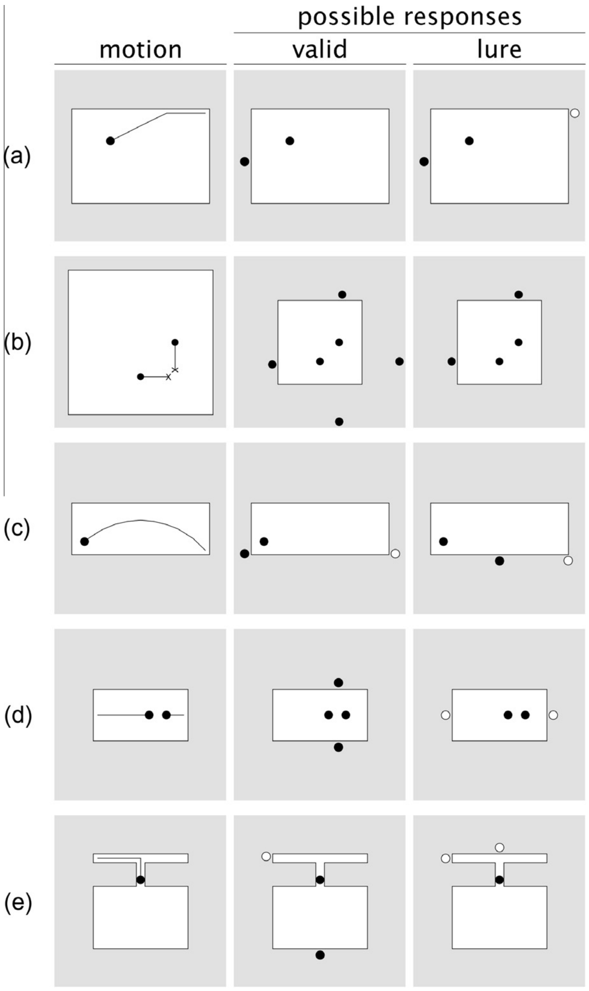

Fig. 10. The (a) wall, (b) centering, (c) curved, (d) split, and (e) lane motions, along with valid explanations for those motions, and simple-but-flawed "lure" responses. The circles within the arenas represent the initial locations of the particles, and the circles outside the arenas represent attractors (white circles) and repellers (black circles). The "x"s for the centering motion denote the point in the particle motion at which participants were informed that the camera had malfunctioned.

corresponded to a rotated version of the explanation that the participant had provided in the discovery phase. The rotation, which was either 90° or 270°, was intended to make it more likely that participants would respond by making a forward inference rather than by recalling and copying the motion from the relevant discovery scene. The other prediction trial for each focal discovery scene corresponded to rotated versions of the valid explanations depicted in Fig. 10.

4.1.2.3. Evaluation phase. On the focal trials in the evaluation phase, participants chose between their own explanations and the simple, valid explanations in Fig. 10.

4.2. Results

Our primary finding is that participants often provided explanations during the discovery phase that were inconsistent with their forward inferences during the prediction phase. Fig. 11 shows the particle motions that the participants predicted when provided with their own

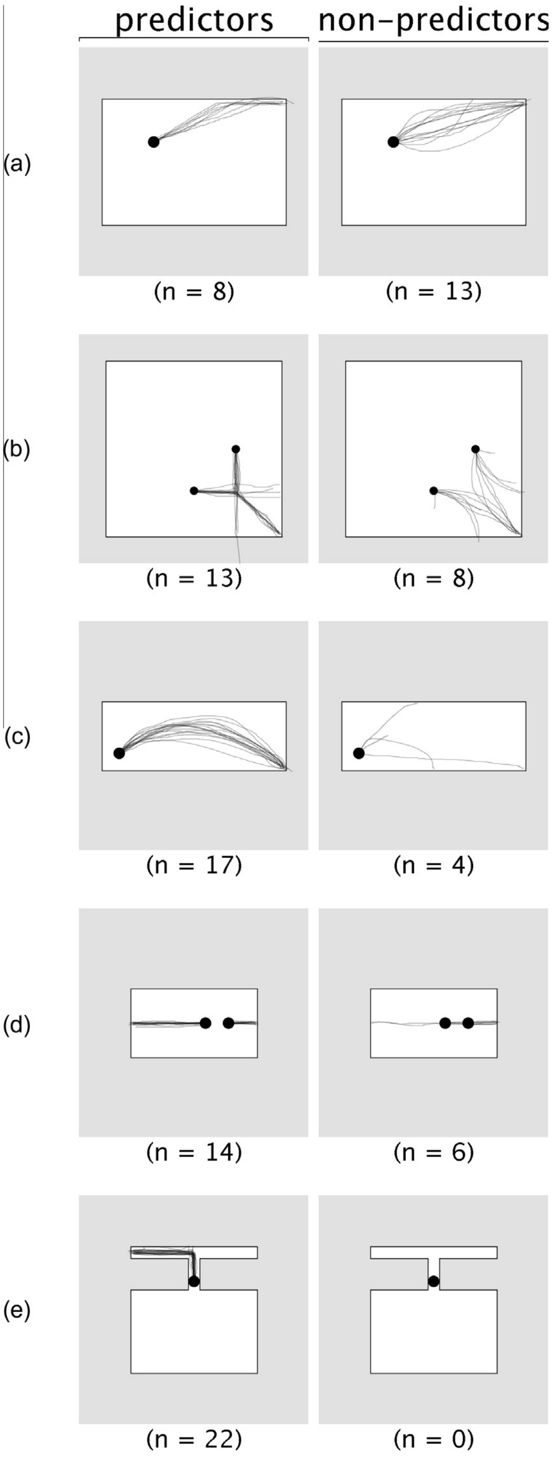

Fig. 11. Particle motions predicted by participants when given their own explanations for the (a) wall, (b) centering, (c) curved, (d) split, and (e) lane motions. Participants were classified as "predictors" when the predicted particle motions were consistent with the actual particle motions from the discovery scenes.

explanations for the particle motions in the discovery phase. Note that many of these predicted motions were inconsistent with the actual motion that was supposedly being explained.

4.2.1. Discovery phase

Fig. 12 shows the most common explanations for each focal discovery scene. Note that lure responses were somewhat common for most of the discovery scenes. This result suggests that some participants provided explanations that they would have recognized as erroneous had they checked them by carrying out a forward inference. Fig. 12 also shows that, as expected, participants rarely provided one of the simplest possible explanations that actually explained the particle motion (see the “simplest valid” explanations).

Among the other explanations, the most common explanations were “modified lure” explanations. These explanations resembled the “lure” responses but posited additional hidden objects. The additional hidden objects often represented efforts to address the problems of the standard lure response. For the split motion, for example, the most common response posited a single attractor at the right and two attractors at the left; the additional attractor at the left might explain why the leftmost particle traveled to the left. The responses that are not represented in Fig. 12 were either explanations that were unusual or difficult to categorize or trials on which the participant did not find an explanation. There were 5 uncategorized responses for the wall motion, 5 for the centering motion, 10 for the curved motion, 2 for the split motion, and 3 for the lane motion. There were two “no explanation” responses for the wall motion, two for the centering motion, two for the curved motion, three for the split motion, and one for the lane motion.

4.2.2. Prediction phase

Fig. 13 shows the predicted particle motions that participants provided when given their own explanations from the discovery phase, conditional on the type of explanation that was provided. Although many of the predicted motions resembled the actual motions from the discovery phase, many of the predicted motions did not. This was especially apparent among the participants who generated the wall-, centering-, and split-motion lures (see the “lure” predictions in rows “a”, “b”, and “d”). In particular, five out of the six participants who generated the wall-motion lure predicted that it would produce a curved particle motion (in contrast to the straight-line motion that the lure was supposed to explain). Likewise, five out of the thirteen participants who generated the centering-motion lure predicted that it would produce curved particle motions (in contrast to the observed straight-line motions). Finally, all of the four participants who generated the split-motion lure predicted that both of the particles would travel to the same side of the arena (in contrast to the observed motion, in which the particles traveled to different sides of the arena). The “explanations” that these participants gave were therefore erroneous, even by the participants’ own accounts. It was only when these participants were asked to make a forward inference given their own explanations that they belatedly recognized the problems with their explanations. This result suggests that at least some participants neglected to check the validity of the explanations that they provided in the discovery phase, and only recognized the problems with these explanations after performing a forward inference in the prediction phase.

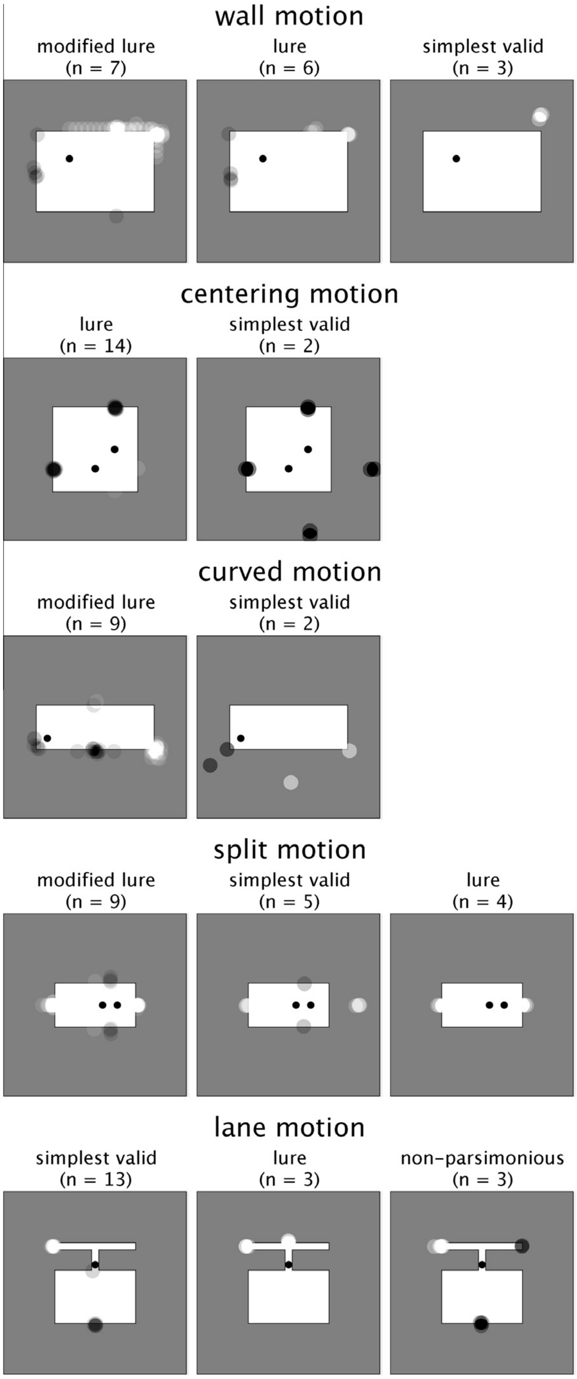

Fig. 12. The most common explanations for the wall, centering, curved, split, and lane motions. Note that “lure” responses were somewhat common: they were often offered for the centering motion and occasionally offered for the wall, split, and lane motions. The “modified” lures resembled the lures but posited additional attractors and repellers.

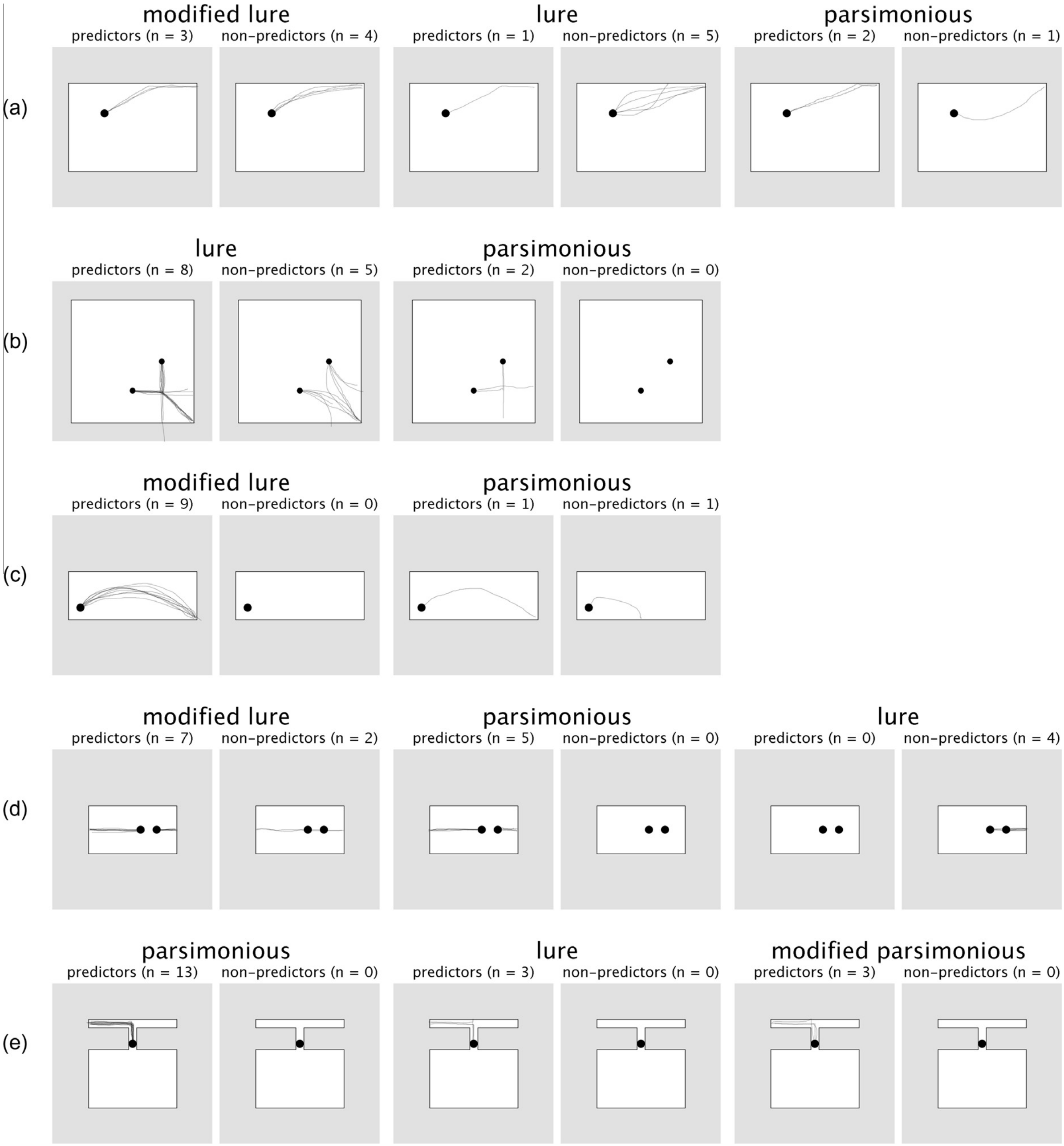

Fig. 13. The particle motions that the participants predicted given their own explanations for the (a) wall, (b) centering, (c) curved, (d) split, and (e) lane motions. The predictions are displayed as a function of the discovery scene, the type of explanation provided, and consistency of the predicted motion with the actual motion from the corresponding discovery scene ("predictors" vs. "non-predictors"). Note that a substantial minority of participants were classified as non-predictors.

Participants were also asked to make predictions given the valid explanations from Fig. 10. As expected, these predicted particle motions generally resembled the actual particle motions from the discovery phase (see Fig. 14 for details). As in the analyses for the previous experiments, we classified participants as "predictors" or "non-predictors" on the basis of these responses.

4.2.3. Evaluation phase

As expected, predictors were much more likely to endorse one of the simplest possible valid explanations in the evaluation phase than to generate one in the discovery phase (see Fig. 14), $p < .001$ by Fisher's exact test. This finding is consistent with the conclusion from Experiments 1 and 2 that participants often failed to consider the best explanations during the discovery phase.

4.3. Discussion

Many of the explanations offered for the particle motions in the discovery phase were flawed, and many of the participants who provided those flawed explanations were capable of recognizing the problems with the explanations (as evidenced by their subsequent inferences in the prediction phase). So why did participants provide these explanations anyway? One possibility is that participants knew that the explanations were flawed but provided them anyway. This possibility seems unlikely, however. To begin with, participants were not required to generate an explanation: they were encouraged to use the ''no explanation'' button whenever they could not find an explanation, and 10 out of the 23 participants used this option at least once during the experiment. Additional evidence against this possibility comes from a debriefing question which asked participants whether they ever provided an ''inadequate or incorrect'' explanation, and, if so, whether they knew that the explanation was inadequate or incorrect at the time that they provided it. Fourteen participants reported providing an invalid explanation, and only two of those participants claimed that they knew that the explanation was invalid at the time that they provided it. A more likely possibility is that participants simply failed to recognize the problems with the explanations at the time that they provided them. These participants presumably generated the explanations heuristically and then failed to check those explanations by performing the corresponding forward inference. Our data therefore suggest that people sometimes make backward inferences without relying on forward inference at all, which implies that the processes that support forward and backward inference are more distinct than the inverse reasoning account allows.

5. General discussion

Our experiments showed that the psychological processes that support discovery, prediction, and evaluation are less closely intertwined than the inverse reasoning account implies. Many participants in Experiment 1, for example, enthusiastically endorsed parsimonious explanations in the evaluation phase even while failing to generate those explanations in the discovery phase. While this finding would not be surprising if the parsimonious explanations were complicated or obscure, Experiment 2 suggested that the parsimonious explanations were relatively accessible: participants were easily able to discover them when heuristic explanations were not available. Experiment 2 therefore suggests that participants only failed to discover the parsimonious explanations because they rarely considered more than a few possible explanations of the particle motions. Experiment 3 exposed an even more fundamental dissociation between discovery and prediction by showing that participants sometimes ''discovered'' explanations that did not actually predict the to-be-explained particle motions. Taken together, these results pose a serious challenge for the inverse reasoning approach to object discovery.

5.1. The computational- and process-level inverse reasoning accounts

Inverse reasoning can be formulated as either a computational-level account about which explanations people will generate or as a process-level account that characterizes the psychological processes that generate these explanations. Our experiments show that neither of these formulations provides a complete psychological account of backward inference. The computational-level inverse reasoning account predicts that the reasoner will identify the best explanation of the available observations. Experiments 1 and 2, however, show that people sometimes fail to consider the best explanations for the observations. Instead, participants in these experiments considered a surprisingly small set of possible explanations for the particle motions, and their decisions about which explanations to consider strongly influenced which explanations they ultimately provided. A complete account of object discovery will therefore need to address how participants decide which explanations to consider.

The inverse reasoning account can also be formulated as a process-level account, and rational process models provide some prominent examples of this approach (e.g., Sanborn et al., 2010; see also Brown & Steyvers, 2009; Ullman et al., 2010). This process-level formulation, however, also seems inconsistent with our data. The core claim of the process-level account is that forward and backward inference are closely connected, but the results from Experiment 3 show that forward and backward inference are supported by distinct psychological processes in at least some settings. The finding that people sometimes generate invalid explanations challenges both computational- and process-level formulations of the inverse reasoning account.

There is, however, at least one version of the inverse reasoning account that is compatible with Experiment 3. It is possible that there are multiple cognitive systems that carry out forward inference. Inverse reasoning engages one of these systems, and a different system is engaged when people are asked directly to solve forward inference tasks. If these two systems operate according to different principles, then forward inference and backward inference tasks may sometimes produce incompatible results even if backward inference is carried out by inverse reasoning.[7]

There is some precedent for the idea that there are multiple systems for forward inference. When people are shown a ball exiting from a curved tube, the predictions they make about the ball's subsequent trajectory can vary depend on whether they are asked to sketch this trajectory or to choose among several simulated trajectories (Kaiser, Proffitt, & Anderson, 1985). Although it is plausible that conceptual and perceptual-motor tasks engage different systems for forward inference, all of the tasks in our experiments were high-level reasoning tasks, and it seems less plausible that these tasks engaged different systems for forward inference. In addition, the inverse reasoning account is appealing in part because it provides a unified

[7] We thank an anonymous reviewer for identifying this possibility.

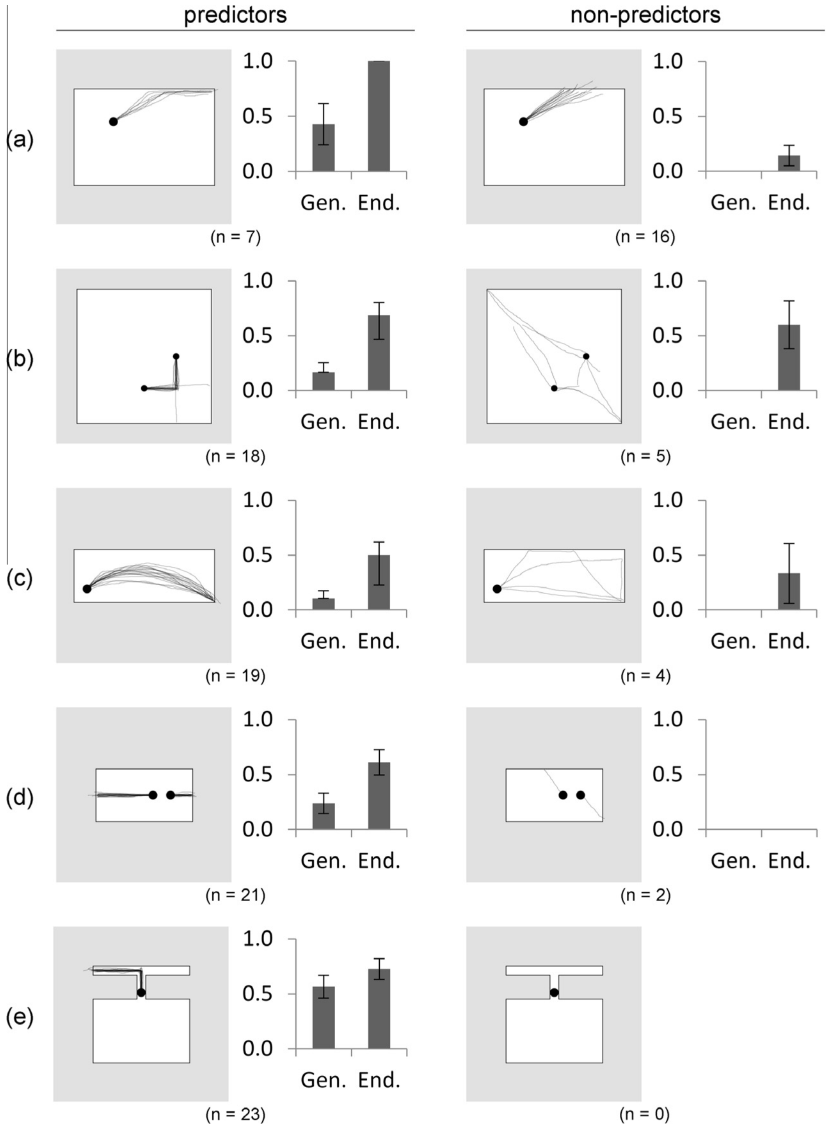

Fig. 14. Inferences about the (a) wall-, (b) centering-, and (c) curved-, (d) split-, and (e) lane-motion scenes in Experiment 3. For each scene, participants were classified as "predictors" or "non-predictors" depending on whether they predicted that the valid explanation from Fig. 10 would produce the actual motion observed in the discovery phase. The bar graphs show the proportion of participants who generated and endorsed one of the simplest possible valid explanations. Gen. = generated; End. = endorsed.

account of forward and backward inference. Adjusting the account by invoking multiple systems for forward inference seems no more parsimonious than simply postulating that forward and backward inference are sometimes carried out by different systems.

Our preferred way to accommodate the results of Experiment 3 is depicted in Fig. 15. The figure presents an account of backward inference that reserves a role for inverse reasoning (shown as the dashed rectangle in Fig. 15), because inverse reasoning is probably involved in many backward inferences. Recall that inferences on the evaluation trials, for example, were largely consistent with the inverse reasoning account. Fig. 15, however, suggests that inverse reasoning plays a more limited role in backward inference than one might initially expect. Under this view, decisions about which explanations to consider (see the arrow from H to H') are among the most important steps in discovering explanations: given that our participants often seemed to consider no more than a few possible explanations of a particle motion, most of the inferential "work" is done by the psychological processes that generate those explanations rather than by the psychological processes that select among those explanations. The modified view of backward inference also allows that people may evaluate potential explanations without performing forward inferences (i.e., without using inverse reasoning). In particular, this view suggests that people sometimes generate explanations without performing any forward inferences at all (see the bold paths in Fig. 15). This sort of situation might arise when people

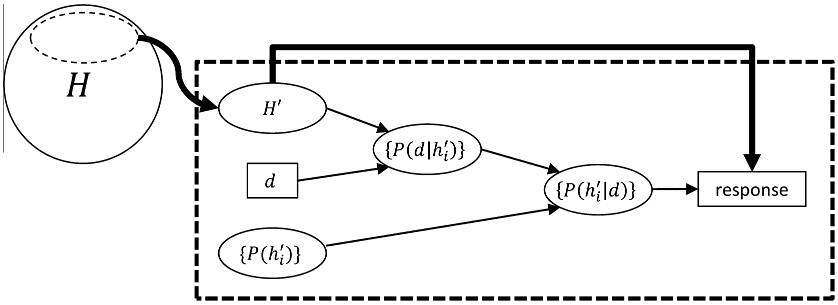

Fig. 15. An outline of the psychological processes that support object discovery. The dashed rectangle shows the scope of the inverse reasoning account, and the bold paths represent inferential processes that are not explained by inverse reasoning. d = the data; H = the full hypothesis space; H' = a subset of the full hypothesis space; $\{P(h')\}$, $\{P(d|h')\}$, and $\{P(h'|d)\}$ = collections containing the prior probabilities, likelihoods, and estimated posterior probabilities of each hypothesis in H'.

accept on faith that certain heuristics for generating possible explanations will produce valid explanations. This represents a reasonable strategy so long as people understand when the heuristics produce valid explanations, but the results of Experiment 3 suggest that people sometimes apply these heuristics in situations where they are not appropriate.

5.2. The problem-solving account of discovery

Although the view of object discovery in Fig. 15 is not entirely compatible with the inverse reasoning account, it can be accommodated by the problem-solving account of scientific discovery. The problem-solving account characterizes scientific discovery as a search through a space of possible hypotheses (Klahr & Dunbar, 1988; Langley et al., 1987). Some search algorithms can be viewed as process-level implementations of the inverse-reasoning account, and algorithms of this kind can be viewed as rational process models (Griffiths et al., 2012). The problem-solving approach, however, also allows for search strategies that are not directly related to inverse reasoning, and that therefore lie outside the class of rational process models. This flexibility allows the problem-solving approach to explain some experimental findings that are inconsistent with rational process models and inverse reasoning. For example, consider the particle motions in Experiment 3 that involved two particles (the “centering” and “split” motions). Faced with these motions, some participants provided “lure” responses that naively combined an explanation for the motion of one of these particles with an explanation for the motion of the other particle. This sort of mistake is inconsistent with inverse reasoning, but is easily explained as the result of a “divide-and-conquer” search algorithm.

When formulated in the most general terms, the problem-solving account seems indisputable but not especially informative. Any process for generating an explanation can be characterized as a search of some kind, and the real question is whether some of the specific search strategies described in the problem-solving literature are able to account for our data. To our knowledge, there is no existing computational model of problem solving that can be applied to our experiments right “out of the box,” but the rest of this section discusses several specific proposals from the problem solving literature that are relevant to our work.

A key element of any search algorithm is a stopping criterion that specifies when the algorithm should terminate the search and return the best solution found so far. The literature on problem-solving proposes that people often “satisfice” and terminate their search after finding a solution that seems acceptable even if it is not optimal (Simon, 1955). This proposal is consistent with our finding that participants often generated explanations that were valid in the sense that they accounted for the available observations, but not as parsimonious as they might have been. Our results therefore suggest that our participants often satisficed and terminated their search for explanations after generating a single valid candidate.

The problem-solving literature includes several specific proposals about mechanisms for searching through a space of hypotheses. According to Simon et al. (1981), the three most common mechanisms are generate-and-test, heuristic search, and means-ends analysis, and we consider each one in turn. The generate-and-test approach repeatedly generates candidate explanations then checks them to see whether they explain the available data. This process of checking an explanation relies on forward inference, and the generate-and-test approach can therefore be viewed as a form of inverse reasoning. For example, if each explanation is either consistent or inconsistent with the data, and if explanations are generated in order of descending prior probability, then the first explanation accepted by a generate-and-test strategy will be the explanation with highest posterior probability. Given the close relationships between generate-and-test and inverse reasoning, it seems unlikely that a generate-and-test approach will provide a complete account of our data.

Heuristic search involves proposing new hypotheses by modifying hypotheses that have already been considered. In the context of our experiments, a participant might begin with a focal hypothesis that does not posit any attractors or repellers. In this context, heuristic search might involve searching for a way to reduce the discrepancy between the focal hypothesis’s predicted motion and the observed motion. Our participants often seemed to follow this strategy. For example, consider the steps

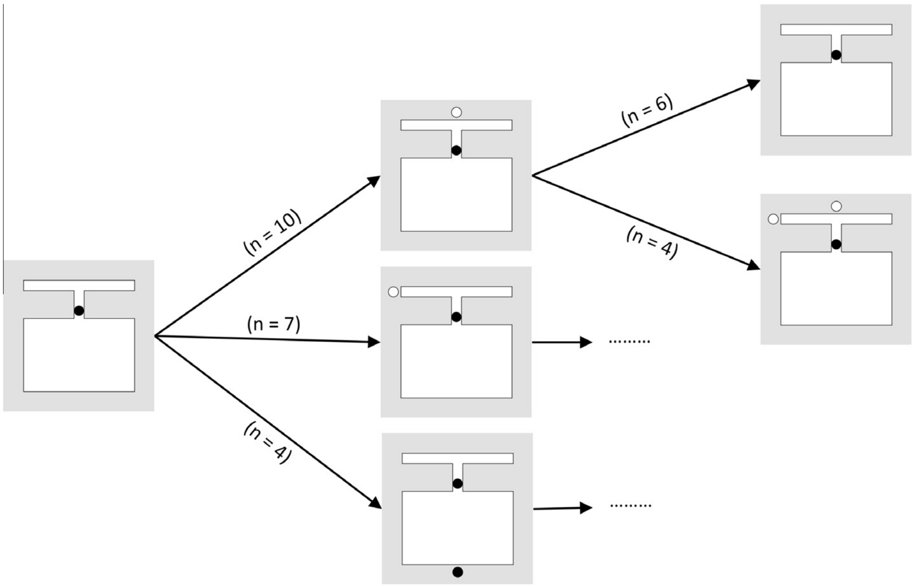

Fig. 16. Initial steps taken to construct explanations for the lane motion from Experiment 3.

taken by our participants to construct explanations for the lane motion from Experiment 3 (see Fig. 16). As a first step in constructing an explanation, many of our participants placed an attractor near the top of the arena (see the uppermost arena in the second column of Fig. 16). Although the existence of an attractor in this location would explain the initial motion of the particle, placing an attractor in that location represents a false start with respect to constructing an explanation for the entire particle motion. Recognizing this, many participants who placed this attractor subsequently "backtracked" by erasing the particle and starting the search anew (see the uppermost arena in the third column of Fig. 16).[8] In any case, the prevalence of this false start suggests that many participants rely on "local" heuristic search strategies to construct explanations.

Means-ends analysis is a special case of heuristic search in which the search is guided by introducing subgoals whenever the current goal cannot be achieved in a single step. It seems likely that our participants relied on means-ends analysis in cases in which the observed particle motion could be divided into several pieces. For example, responses to the lane-motion scene suggest that participants often viewed explaining the upward and leftward motions of the particle as two independent sub-problems. More generally, means-ends analysis seems applicable to any of the discovery scenes in which there were multiple particles in motion (e.g., the "centering" and "split" motions of Experiment 3) or in which the motion of a single particle could be subdivided into distinct phases (e.g., the "wall" motions).

Although problem-solving strategies such as heuristic search and means-ends analysis seem broadly consistent with our data, developing a computational framework that can account for the variety of responses observed across our experiments appears to be a substantial challenge. A comprehensive account of our data will probably need to incorporate ideas that go beyond general-purpose methods such as heuristic search and means-end analysis. Experts often make extensive use of domain-specific knowledge when solving problems (e.g., Chi, Feltovich, & Glaser, 1981; Larkin et al., 1980; Patel & Groen, 1986), and our participants may also have relied on domain-specific knowledge. For example, the "orbiting" explanation for the curved motion stimulus in Fig. 8 may have been inspired by prior knowledge about systems in which one object (e.g. the earth) orbits another (e.g. the sun). We therefore believe that developing a problem-solving account of our results may be possible, but is by no means simple.

5.3. How can the successes and failures of the inverse reasoning account be reconciled?

Although we have argued that the inverse reasoning account has limitations as a psychological model of backward inference, the inverse reasoning account often explains people's inferences well. How are we to reconcile the successes and failures of the inverse reasoning account? A partial answer is that the problem of hypothesis generation is simply more acute in some inferential tasks than in others. Consider the comparison between our object-discovery task and the task where participants are asked to infer the mass ratio of two colliding objects (e.g., Sanborn et al., 2013). Although there are an infinite number of possible explanations in both tasks, the problem of generating the best possible explanations is considerably less daunting in the object-collision task. When

[8] Of the four participants who did not backtrack, three provided explanations corresponding to the second arena in the third column of Fig. 16 (i.e., the "lure" response). The other participant placed three attractors at the left end of the "lane".

faced with the mass-ratio task, for example, the reasoner might imagine what sort of collision would have occurred for a particular mass ratio. By comparing that imagined collision to the actual collision, the reasoner would be able to decide whether the given mass ratio is too small or too large, allowing the reasoner to modify the proposed mass ratio accordingly. The iterative application of this hill-climbing procedure would allow the reasoner to quickly discover the mass ratio that best explains the observed object collision. In our object-discovery task, in contrast, the prospects for finding a procedure that efficiently identifies the best explanation seem dim.

Expertise and experience undoubtedly play a further role in explaining the successes of the inverse reasoning account. Many of the most successful applications of the inverse reasoning account involve explaining inferences that people make in their everyday lives. Perhaps extensive experience with these tasks has taught people which explanations should be considered in any given situation, much as expert problem solvers know which problem-solving strategies are appropriate to which problems (e.g., Chi et al., 1981; Patel & Groen, 1986). Some inferences that resemble inverse reasoning might therefore be attributed to the reasoner's considerable experience and expertise with those tasks rather than to inverse reasoning in and of itself. This sort of account—wherein forward and backward inference are performed by separate psychological processes, but in which extensive training allows the processes that support backward inference to approximate inverse reasoning—is often proposed to explain motor planning and control (e.g., Davidson & Wolpert, 2005; Flanagan, Vetter, Johansson, & Wolpert, 2003; Jordan & Rumelhart, 1992; Kawato, 1999).

5.4. Conclusion

The inverse reasoning approach has been widely used to account for inductive reasoning, but our results suggest that this approach is incomplete at best as a psychological account of object discovery. The approach predicts that discovery, evaluation, and prediction should be closely related and mutually consistent, but our experiments demonstrate that these inferences are often incompatible. In particular, we found that people often endorsed explanations that they were unable to generate themselves. This inconsistency arises because generating an explanation appears to be a much more challenging problem than assessing the merits of several pre-specified explanations.

Our findings underline the importance of understanding how people search the hypothesis space of possible explanations. Our experimental setting focused on inferences about object dynamics, and we found that participants often relied on domain-general strategies such as heuristic search and on domain-specific strategies such as placing unobserved objects directly in line with the motion of observed objects. Different strategies may be more or less useful in different settings, and developing a comprehensive account of the psychological processes that support object discovery will require multiple settings to be explored in detail.

References

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

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.

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

Brown, S. D., & Steyvers, M. (2009). Detecting and predicting changes. *Cognitive Psychology, 58*(1), 49–67.

Chater, N., & Manning, C. D. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences, 10*(7), 335–344.

Chi, M. T., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*, 121–152.

Churchland, P. M., & Hooker, C. A. (Eds.). (1985). *Images of science: Essays on realism and empiricism*. University of Chicago Press.

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

Csibra, G., & Volein, A. (2008). Infants can infer the presence of hidden objects from referential gaze information. *British Journal of Developmental Psychology, 26*(1), 1–11.

Davidson, P. R., & Wolpert, D. M. (2005). Widespread access to predictive models in the motor system: A short review. *Journal of Neural Engineering, 2*(3), 313–319.

diSessa, A. A. (1993). Toward an epistemology of physics. *Cognition and Instruction, 10*(2–3), 105–225.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415*, 429–433.

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.

Fernbach, P. M., Darlow, A., & Sloman, S. A. (2011). Asymmetries in predictive and diagnostic reasoning. *Journal of Experimental Psychology: General, 140*(2), 168–185.

Fernbach, P. M., & Sloman, S. A. (2009). Causal learning with local computations. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 35*(3), 678–693.

Fienberg, S. E. (2006). When did Bayesian inference become "Bayesian"? *Bayesian Analysis, 1*(1), 1–40.

Fischer, P., & Zytkow, J. M. (1992). Incremental generation and exploration of hidden structure. In J. M. Żytkow (Ed.), *Proceedings of the ML-92 Workshop on Machine Discovery* (pp. 103–110). Wichita, KS: National Institute for Aviation Research.

Flanagan, J. R., Vetter, P., Johansson, R. S., & Wolpert, D. M. (2003). Prediction precedes control in motor learning. *Current Biology, 13*(2), 146–150.

Gerstenberg, T., Goodman, N., Lagnado, D. A., & Tenenbaum, J. B. (2012). Noisy Newtons: Unifying process and dependency accounts of causal attribution. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 378–383). Austin, TX: Cognitive Science Society.

Gilden, D. L., & Proffitt, D. R. (1989). Understanding collision dynamics. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 15*(2), 372–383.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review, 114*(2), 211–244.

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

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.

Howson, C., & Urbach, P. (1996). *Scientific reasoning: The Bayesian approach* (3rd ed.). Open Court Publishing Co. (Original work published 1989).

Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge, UK: Cambridge University Press.

Jordan, M. I., & Rumelhart, D. E. (1992). Forward models: Supervised learning with a distal teacher. *Cognitive Science, 16*(3), 307–354.

Kaiser, M. K., Proffitt, D. R., & Anderson, K. (1985). Judgments of natural and anomalous trajectories in the presence and absence of motion. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 11*(4), 795.

Kawato, M. (1999). Internal models for motor control and trajectory planning. *Current Opinion in Neurobiology, 9*(6), 718–727.

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review, 116*(1), 20–58.

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science, 12*, 1–48.

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

Koriat, A., Lichtenstein, S., & Fischoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6*(2), 107–118.

Langley, P., Simon, H. A., Bradshaw, G. L., & Zytkow, J. M. (1987). *Scientific discovery: Computational explorations of the creative processes*. MIT Press.

Larkin, J., McDermott, J., Simon, D. P., & Simon, H. A. (1980). Expert and novice performance in solving physics problems. *Science, 208*, 1335–1342.

Lipton, P. (2004). *Inference to the best explanation*. Psychology Press.

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology, 55*(3), 232–257.

Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review, 115*(4), 955–984.

Marcus, G. F., & Davis, E. (2013). How robust are probabilistic models of higher-level cognition? *Psychological Science, 24*(12), 2351–2360.

Marr, D. (1982). *Vision*. San Francisco, CA: W.H. Freeman.

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

McCloskey, M., Caramazza, A., & Green, B. (1980). Curvilinear motion in the absence of external forces: Naive beliefs about the motion of objects. *Science, 210*, 1139–1141.

Patel, V. L., & Groen, G. J. (1986). Knowledge based solution strategies in medical reasoning. *Cognitive Science, 10*, 91–116.

Polya, G. (1954). *Mathematics and plausible reasoning. Patterns of plausible inference* (Vol. II). Princeton University Press.

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.

Sanborn, A. N., Mansinghka, V. K., & Griffiths, T. L. (2013). Reconciling intuitive physics and Newtonian mechanics for colliding objects. *Psychological Review, 120*(2), 411–437.

Saxe, R., Tenenbaum, J. B., & Carey, S. (2005). Secret agents: Inferences about hidden causes by 10- and 12-month-old infants. *Psychological Science, 16*(12), 955–1001.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics, 69*(1), 99–118.

Simon, H. A., Langley, P. W., & Bradshaw, G. L. (1981). Scientific discovery as problem solving. *Synthese, 47*(1), 1–27.

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

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

Tarantola, A. (2006). Popper, Bayes and the inverse problem. *Nature Physics, 2*(8), 492–494.

Teglás, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J. B., & Bonatti, L. L. (2011). Pure reasoning in 12-month-old infants as probabilistic inference. *Science, 332*, 1054–1059.

Thomas, R. P., Dougherty, M. R., Sprenger, A. M., & Harbison, J. I. (2008). Diagnostic hypothesis generation and human judgment. *Psychological Review, 115*(1), 155–185.

Todd, J. T., & Warren, W. H. (1982). Visual perception of relative mass in dynamic events. *Perception, 11*(3), 325–335.

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review, 101*(4), 547–567.

Ullman, T. D., Goodman, N. D., & Tenenbaum, J. B. (2010). Theory acquisition as stochastic search. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 2840–2845). Austin, TX: Cognitive Science Society.

van Fraassen, B. C. (1980). *The scientific image*. Oxford University Press.

Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General, 121*(1), 222–236.

Wolpert, D. M., & Kawato, M. (1998). Multiple paired forward and inverse models for motor control. *Neural Networks, 11*(7), 1317–1329.

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