Cognition 199 (2020) 104238

Contents lists available at ScienceDirect

Cognition

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

Incidental binding between predictive relations

Anna Leshinskaya*, Mira Bajaj, Sharon L. Thompson-Schill

Department of Psychology, University of Pennsylvania, United States of America

ARTICLE INFO

Keywords:
Predictive learning
Structure learning
Statistical learning

ABSTRACT

Knowledge of predictive relations is a core aspect of learning. Beyond individual relations, we also represent intuitive theories of the world, which include interrelated sets of relations. We asked whether individual predictive relations learned incidentally in the same context become associatively bound and whether they spontaneously influence later learning. Participants performed a cover task while watching three sequences of events. Each sequence contained the same set of events, but differed in how the events related to each other. The first two sequences each had two strong predictive relations (R1 & R2, and R3 & R4). The third contained either a consistent pairing of relations (R1 & R2) or an inconsistent pairing (R1 & R3). We found that participants' learning of the individual relations in the third sequence was affected by pairing consistency, suggesting the mind associates relations to each other as part of the intrinsic way it learns about the world. This was despite participants' minimal ability to verbally describe most of the relations they had learned. Thus, participants spontaneously developed the expectation that pairs of relations should cohere, and this affected their ability to learn new evidence. Such associative binding of relational information may help us build intuitive theories.

1. Introduction

Part of what makes human cognition so sophisticated is that we represent not a catalog of sensory facts, but rather, coherent world models (*theories*) that explain and predict observations (Carey, 2009; Gelman & Wellman, 1991; Gopnik, 1996; Gopnik & Meltzoff, 1997; Gopnik & Wellman, 1994; Keil, Smith, Simons, & Levin, 1998; Kemp, Tenenbaum, Niyogi, & Griffiths, 2010; Lombrozo, 2009; Tenenbaum, Kemp, Griffiths, & Goodman, 2011). In a canonical example, our theory of mind explains people's actions by relating what they see to what they know, and what they know and desire to what they do (Baker, Saxe, & Tenenbaum, 2009, 2011; Dennett, 1987; Premack & Woodruff, 1978).

Much other knowledge—about personality traits, technology, biology, cooking—also has a theory-like character, but is unlikely all innate, raising the question of how it could be learned (Schulz, Goodman, Tenenbaum, & Jenkins, 2008). The pervasiveness of theories such as these makes it likely that our minds are equipped and predisposed to build them. Here, we describe an automatic process that could form part of such a mechanism. Specifically, we claim that in the course of spontaneous associative learning, the human mind is already inclined to build knowledge structures which have a theory-like character.

One distinguishing feature of theories, as opposed to individual relations, is that they are coherent, interrelated *sets* of relations (Gopnik & Meltzoff, 1997). In a theory of mind, an agent will believe what he perceives (relation 1) *and* act on what he desires (relation 2). The holder of the theory believes that if she observes relation 1, that relation 2 should also hold. But prior to having the theory, how would she know that these individual dependencies hang together?

Prior work demonstrates that both adults and children can learn how multiple predictive relations covary on the basis of statistical evidence, and use this to reason about new scenarios. Schulz and colleagues (Schulz et al., 2008) found that pre-schoolers infer that multiple causal relations about the behaviors of novel blocks will hang together in the future if they have in the past. If a red block 'activates' a blue block, and a blue block activates a yellow block, children readily infer that a novel block activated by a red block (acting as if blue) will also activate the yellow one: they generalized the *pairings* of relations. Adults do the same during explicit causal reasoning (Waldmann, Meder, Von Sydow, & Hagmayer, 2010). Gershman (2017) showed that adults rationally use context variation to guide such inferences: if multiple lower-order relations (e.g., about which of several foods are pleasant vs. aversive) vary by context, they expect such relations to pattern together consistently in new contexts, when asked to reason about them explicitly.

Reasoning of this sort is rational and adaptive. Here we wondered whether co-variation of relations will affect learning itself. In other

* Corresponding author at: Department of Psychology, University of Pennsylvania, 425 S. University Ave, Stephen A. Levin Bldg., Philadelphia, PA 19104, United States of America.

E-mail address: aleshinskaya@ucdavis.edu (A. Leshinskaya).

https://doi.org/10.1016/j.cognition.2020.104238
Received 30 October 2018; Received in revised form 12 February 2020; Accepted 12 February 2020
Available online 29 February 2020
0010-0277/ © 2020 Elsevier B.V. All rights reserved.

words, does the very way in which the mind encodes information include the knowledge of which relations predict each other? In prior work, participants were taught relations in an explicit manner, and then asked to reason on their basis (i.e., about what will happen in new situations). Here we tested whether adult participants will spontaneously encode relations among relations during passive observation, and whether this encoding will inadvertently affect how accurately they learn new predictive relations. If so, tracking the co-variation among relations may be an intrinsic part of how the human mind encodes the world.

To test this idea, we presented participants with four individual predictive relations among sequentially presented events, where pairs of these relations ('relational sets') co-varied across two contexts. In a critical third context, we measured how well participants could learn two similar relations which were paired either consistently, or inconsistently. During the task, there was no demand or benefit to reasoning about pairings of relations; we measured only how well participants could learn each relation individually. However, if the covariation among relations is an intrinsic part of associative learning, then the consistency of their pairings should affect learning, even when this is inadvertent and produces errors. In other words, we propose that the binding of relations into coherent sets might operate similarly to how we spontaneously learn other observed, predictive statistics of the environment (Reber, 1989; Saffran, Aslin, & Newport, 1996)—but at a higher order level, at which relations become associated with other relations.

To test this, it is essential to vary predictive relations while controlling for the individual events involved in them. Imagine that one relation is that flipping a light switch results in the light turning on, and a second relation in the same context is that pressing a button causes a sound. In a different context, learners might anticipate the second rule if they observe the first, but this could happen because they are anticipating the sound to occur, regardless of whether it is related to the button. Thus they could have simply associated the component events, not the relations themselves. To avoid this, one must use multiple contexts in which the same events occur with equal frequency, but are *related* in different ways. In the present experiment, we specifically target the associability of relations themselves in this way. We thus address a distinct question from related work on grouping action rules (Collins & Frank, 2013, 2016; Werchan, Collins, Frank, & Amso, 2015).[1] Furthermore, we query observational learning, rather than action-reward learning or stimulus-reward learning, by employing a statistical learning paradigm (Orbán, Fiser, Aslin, & Lengyel, 2008; Saffran et al., 1996; Turk-Browne, Jungé, & Scholl, 2005). This is important because our theories of the world are most often constructed to explain observed events, which often may not have an explicit reward or action associated with them.

In our statistical learning paradigm, we presented participants with continuous sequences of animated events (Fig. 1), which appeared as part of distinct sequences (or 'contexts') distinguished by different objects present in the events. All sequences involved the same eight events, which all appeared with equal frequency. However, the predictive structure among the events varied, so that different subsets of the 8 were predictively related vs unrelated, allowing us to create relational sets. Each sequence contained two predictively related pairs, each involving two events (which we term the 'cause' and an 'effect'), such that the effect almost always followed the cause. The first two ('Training') sequences set up how the individual cause-effect relations themselves were paired. For example, in Sequence A (as shown in Fig. 2), one predictive relation might be that the object tilting is followed reliably by the light flashing (R1), and a second might be the object turning blue predicts the multi-colored stars appearing (R2). The other 4 events appeared equally frequently but were not part of any predictable pairs. In Sequence B, the events that had been part of cause-effect pairs in Sequence A (tilt, light, color change, and stars) became unrelated, while the other 4 events formed two other cause-effect pairs (termed R3 and R4; as shown in Fig. 3). The two training sequences A & B together set up the higher-order structure governing how the individual relations cohered into sets: if R1 holds, R2 should hold; but if R3 holds, R4 should hold, regarding the same set of 8 events. Participants were not told about this structure, only exposed to it. To test whether participants spontaneously encoded this higher order structure among relations, we asked how it affected learning in a third sequence.

In the third, 'Test' sequence, cued by a different object, again the same set of eight events was shown, and two familiar individual cause-effect relations were present. However, the *pairing* of these relations was either Consistent or Inconsistent with the pairing structure previously seen in the two training sequences (following Collins & Frank, 2013). In the example in Fig. 3, the Consistent test sequence exhibited both R1 and R2, individual relations both seen in Sequence A. The Inconsistent sequence exhibited R1 and R3, two relations which were equally familiar, but paired inconsistently—one came from Sequence A, and the other came from Sequence B. We predicted that, despite no instruction to attend to the pairing of relations, participants would have spontaneously attended to their covariation, and because of this, their learning of the individual relations in the test sequence would be affected by their consistency with those pairings.

We measured how well participants learned the individual relations with a forced-choice test, which asked them to choose between clips showing typical predictive relations (cause followed by effect) vs atypical clips (two unrelated events pairs; Fig. 1B). These learning probes did not measure knowledge of how the relations went *together*; it measured only knowledge of the individual relations; that is, which individual events followed which others reliably. However, for the test sequence, we expected that accuracy on these probes should be affected by the way relations were paired (i.e., consistency condition)—that is, if learners spontaneously encoded such relations in terms of higher-order sets. Individual relations learned in an inconsistent set should be harder to learn, because participants had different expectations about their pairings from prior exposure. Thus, although the test always had a right answer, participant's expectations that relations should continue to be paired consistently would impair their accuracy in the inconsistent condition. This would demonstrate that learning itself is inadvertently affected by spontaneously made inferences about how relations cluster together.

2. Method

2.1. Overview of procedures

Participants watched several short (4.5 min) videos depicting sequences of events while performing a cover task, in which they were asked to determine if the event they were seeing was the common or rare alternate (e.g., the blue bubbles were pink 10% of the time). This is depicted in Fig. 1A. Each participant saw three types of sequences in turn: two Training Sequences (A and B) and one Test Sequence (either Consistent or Inconsistent; Fig. 3), which each featured different objects and different predictive statistics among the same set of 8 events. Each sequence was characterized by two strong predictive ('cause-effect') relations: for example, in Sequence A, an object tilt might be nearly always followed by a light flash, and color change nearly always followed by stars, as shown in Fig. 3; neither of these held in Sequence B, although tilt, light flash, color change and bubbles all still took place with equal frequency. By varying only the statistical structure, rather

[1] In these experiments, sets of rules differ in terms of which consequent is more likely to take place. For example, if subjects are taught to press button 1 when a red square appears, and button 2 when a blue circle appears, they will press buttons 1 and 2 more often than in a second context, where they learn to press buttons 3 and 4. Grouping these rules together involves binding the rules, but also binding the two consequent events (buttons 1 & 2).

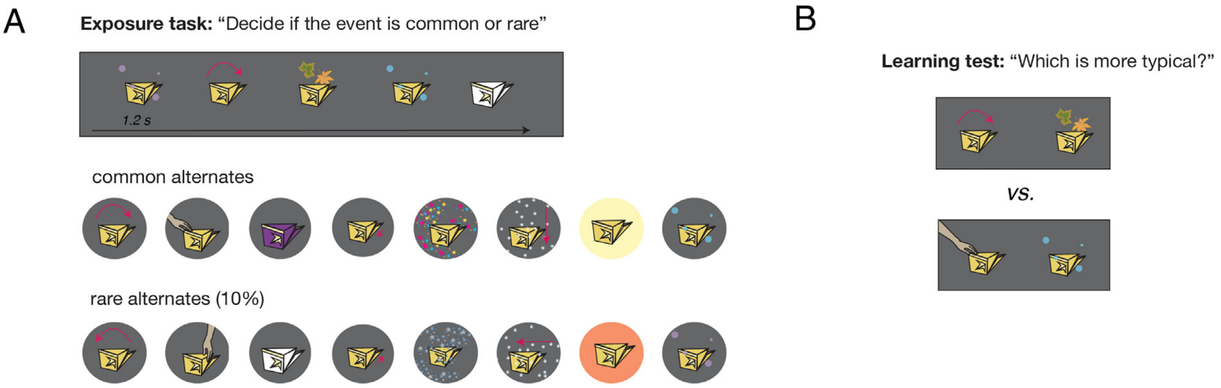
A
Exposure task: "Decide if the event is common or rare"
1.2 s
common alternates
rare alternates (10%)
B
Learning test: "Which is more typical?"
vs.

Fig. 1. A. Illustration of the cover task, in which participants had to decide for each individual event whether it was common or rare. Events appeared in a continuous stream, with the central object continually present as events took place on or around it. Below, images depicting animated event stimuli used in the experiment. The top row shows the 'common' events, and the bottom row shows the 'rare' events. The rare alternates replaced their common versions 10% of the time, at random, for purposes of the cover task. Object based events are the first four pairs on the left, with arrows indicating motion; Ambient events are the next four pairs. **B.** Illustration of the forced-choice test, which presented participants with two, two-frame video clips, and asked them to select which was more typical. This figure is available in larger, PDF format at https://osf.io/5autq/. A shorted demo of the experiment is available for web-view at https://www.sas.upenn.edu/~alesh/images/EXP6T/TaskDemo.html and for download at https://osf.io/jr3u2/.

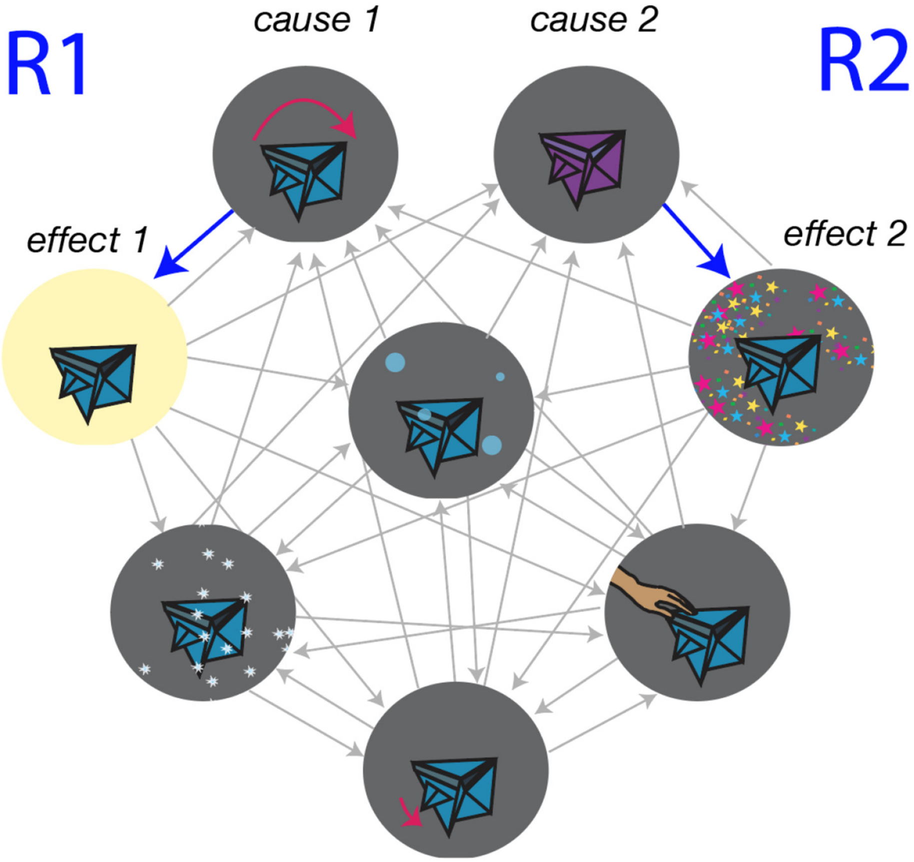

Fig. 2. Graphical depiction of the transition matrix structure governing each sequence, for an example assignment of events. Blue arrows indicate strong (> 90% probability) transitions, while gray arrows indicate equiprobable (~14% probability) transitions, and no arrows indicates a < 5% probability transition. Thus, two strong pairs were exhibited in each sequence, which here are labeled R1 and R2. A larger version of this figure is available at https://osf.io/mc6uz/. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

than which events appeared, we were able to provide participants with information about which predictive relations co-apply in the same context (i.e. co-vary). The third, Test Sequence, either maintained or violated that pairing (between subjects). Our critical dependent measure was a forced-choice test probing knowledge of the specific relations in the sequences (Fig. 1B). On each trial, participants saw two snippets of video, showing either a likely transition (cause followed by effect) or an unlikely one (between two of the unrelated events—in fact, the events that are related in different sequences), and had to select which was more typical. Both individual cause-effect relations were tested separately for each sequence, and the tests were given directly after each sequence was shown. We then measured how well participants had learned the Test Sequence relations relative to their baseline Training Sequence knowledge. We expected that performance would be affected by condition: knowledge of the individual relations in the Test Sequence should be worse in the Inconsistent than in the Consistent condition, relative to baseline learning.

2.2. Participants

We recruited 490 participants using Amazon Mechanical Turk; all were required to have an IP address in the United States and a 95%

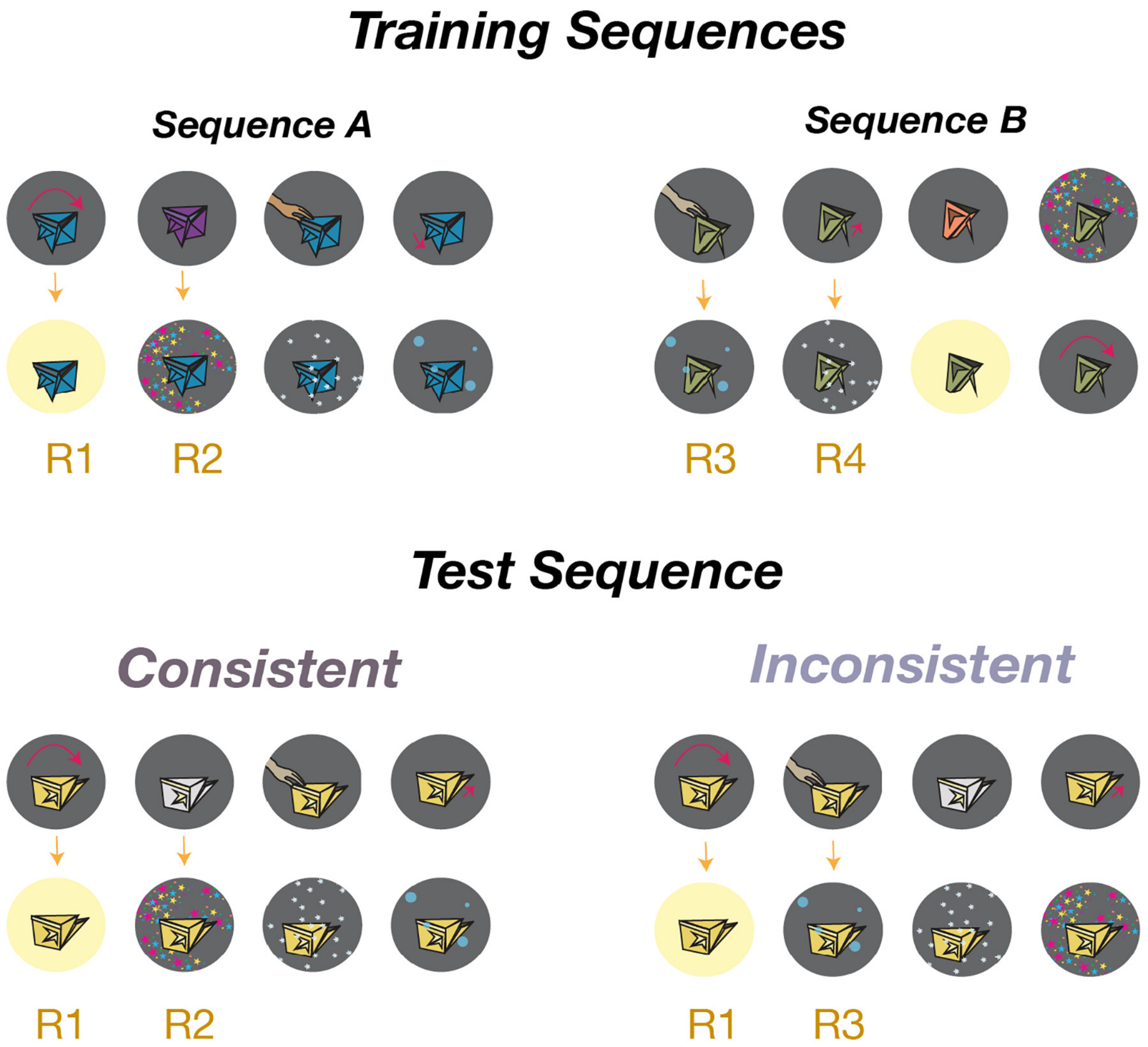

Fig. 3. Illustration of the 4 unique rules and their distribution among the three sequences. All participants saw two training sequences (A and B), each of which exhibited two pairs of relations (R1 & R2, or R3 & R4); specific stimuli were counterbalanced. Half of the participants then saw the Consistent sequence, which exhibited both rules from Sequence A, while the other half saw the Inconsistent Sequence, which exhibited one rule from Sequence A (R1) and one rule from Sequence B (R3). A larger version of this figure is available at https://osf.io/hyex5/.

previous work approval rate. We excluded 37 participants because they had previously participated in a related experiment or repeated this one. An additional 70 were excluded for failing to pass an attention measure (described below). Seven were excluded because of incomplete data or because they reported a technical glitch during the experiment. All excluded participants were replaced to complete a full set of counterbalanced materials (see Stimuli). Of the 376 participants included in the reported analyses, 51% were female (192/376) and the average age was 33.5 (range of 19–68). Procedures were approved by the Institutional Review Board of the University of Pennsylvania, and all participants provided electronic consent. Compensation was $5, with a bonus of up to $5 based on cover task accuracy (described below).

2.3. Sample size determination, effect sizes, and piloting

In an incidental learning paradigm such as this one, a fairly large amount of exposure is typically necessary for participants to learn a complex set of context-varying relations; yet exposure time to the stimuli is also limited by participants' interest and fatigue. This means that our manipulation of participants' experiences—and thus, the relevant comparison—was minimal in each person, as were the number of measurement trials. Since each participant learned two specific predictive relations in each sequence, these could only be tested in a limited number of ways. This means that both the level of exposure to the materials, and the number of measurements of the resulting learning, was very small in each individual.

The approach we took in this work was thus to measure the effect of this minimal intervention in each individual, but with many individuals (rather than collecting many measures in fewer participants). This methodological choice was motivated by these inherent methodological constraints. We first performed a pilot study (with $n = 80$), as described below and in Supplemental Materials, which provided a measure of effect size. Inevitably, the observed and expected effect size was small, but allowed us to determine a necessarily sample size and plan analyses a priori (with a small exception described below).

The final sample size of 376 (following exclusions) was based on two considerations. Firstly, a power analysis using our pilot sample indicated that a sample of 188 participants would be sufficient to obtain 80% power. We did obtain a significant effect for the predicted interaction effect of interest in an initial sample of 188, ($F(1, 186) = 4.75$, $MSE = 0.24$, $p = .031$, partial $\eta^2 = 0.011$). However, we saw order effects in our baseline learning measures that created difficulties of interpretation. To control for order effects, we swapped the order of the relevant measures to fully counterbalance them, and added another 188 participants, which was effective in removing differences between the training stimuli. As we report below, the major finding remained significant.

2.4. Stimuli

Sequences were composed of 8 animated event types (chosen out of a pool of 10 for each participant), most of which are shown in static form Fig. 1A; actual stimuli were GIFs. Each event type had a regular version (top row) and a slightly visually distinct 'oddball' alternate, bottom row of Fig. 1A, for the purposes of the cover task (described below), but sequences were specified over the event types. Four event types were *object-based*: tilting (the entire object rotated 10 degrees), part moving (a detachable part on the object moved/rotated), color changing (the object gradually changed into a different color), or being tapped by a hand. The other six were *ambient*: the background changing color; snowflakes falling; bubbles floating across the scene; confetti swirling; a pair of leaves falling; and a burst of glitter. Each participant saw 8 of the 10 event types, selected randomly. Additionally, *static* "events" in the video streams showed the object still on the gray background and were included to provide intermittent pauses. Each GIF file comprised 12,100 ms frames (total length of 1200 ms), except static (2400 ms).

To create the sequences, events were concatenated into continuous sequences ("videos"). These were generated probabilistically using a weighted walk, where the weights were specified by a pairwise transition matrix that specified the probability of any event following any another. This pairwise transition matrix specified the predictive structure central to our design, shown graphically in Fig. 2 and numerically in Supplemental Tables 1–4. All sequences followed this abstract structure, but varied in how the participant's 8 specific event types were assigned to it. The structure always specified two strongly predictive event pairs, which formed the individual predictive relations: item two in each pair (the "effect") followed item one (the "cause") with a 98% probability. The cause could be followed by a static event with the remaining 2% probability. The effect could repeat with a 1% probability but did not follow any other event. The remaining four events were followed by static, each other, or the cause with a 14% probability. Thus, among the 8 events in a sequence, 2 were causes, 2 were effects, and 4 were weakly predictable (random). All events had equal frequency, as specified by the stationary distribution of the transition matrix.

To ensure each generated sequence was a good reflection of the requested transition matrix, the walks were generated iteratively and verified until they met several criteria: no two events differed in frequency by > 10 instances (~2%), the actual cause-effect transition strength was above 90%, and the other transitions occurred with a probability between 5 and 30%. The obtained transition matrix, averaged across all subjects' walks, is shown in Supplemental Tables 1–4.

After the sequences were generated, the rare alternates were shown instead of the common event versions with a 10% probability, while ensuring that the number of oddballs was equated within 2 instances across the event types.

In the 3 distinct sequences, event types were shown with a different object present (yellow, blue or green; Fig. 3). Sequences also varied in how the 8 specific events were assigned to the same abstract structure—that is, which events were 'causes', 'effects', or random events. Because each sequence had two strong predictive pairs, it had two causes, two effects, and four random events. This created four distinct predictive relations (R1 and R2 for Sequence A, and R3 and R4 for Sequence B). In the third, Test Sequence, we used R1 and R2 in the Consistent condition, and R1 and R3 in the Inconsistent condition (see Fig. 3).

The specific event types assigned to the distinct roles in the four relations were selected for four yoked participants at a time. These assignments were chosen randomly, with the constraint that any of the four types of object-based events (tilt, part-move, color change, or tap) could serve as one of the four causes, and any four of the six ambient events could serve as one of the four effects (with the other two then not shown). An additional constraint was that tilt and part-move could not serve as causes within the same object, as they could be confusable. The rationale for the design overall was to mimic an aspect of the real world, in which the behavior/actions of objects lead to outcomes in the environment, and multiple relations might apply to the same object.

The relations shown in the Test Sequence varied between participants with respect to the relations shown in the Training Sequences (A and B). In the Consistent condition, the Test Sequence had the same cause-effect relations as Sequence A (R1 & R2). In the Inconsistent condition, it exhibited one relation from Sequence A (R1) and one relation from Sequence B (R3). These assignments were thus highly systematic and always followed this abstract structure: i.e., Sequence A always matched the Consistent Test Sequence. For this reason, the specific objects assigned to Sequence A and B and their order of presentation were counterbalanced across conditions (green or blue); The Test Sequence always used the yellow object. For Sequences A and B, three videos were created to be 225, 200, and 200 events in length, each adhering to the sequence properties described earlier. The Test Sequence was shown over two videos, of 225 and 200 events in length.

For the four yoked participants, half had the Consistent Test Sequence and half the Inconsistent Test Sequence. Otherwise, they saw identical materials for the Training Sequences (including the randomly generated walks). Additionally, the event assignments for R2 and R3 were counterbalanced across conditions, within the yoked set: they were exchanged so that if one pair of participants saw color-change–stars for R2 and hand-tap–bubbles for R3, as depicted in Fig. 3, then another pair saw hand-tap–bubbles for R2, and color-change–stars for R3. This was because R2 and R3 are the critical rules differing between conditions in the Test Object (see Fig. 3). Thus, the identity of the events comprising the Test Sequence relations was perfectly counterbalanced across conditions.

2.5. Procedure

Participants were randomly assigned both to an experimental condition (Consistent vs. Inconsistent), and to a counterbalancing set. The experiment was implemented using JavaScript and presented in a web-browser, via the Mechanical Turk interface. Participants could access the experiment during the daytime, between 10 am and 7 pm EST, and had to complete the procedure within 2 h. The average duration of the procedure was 77.84 min.

The participants' task was to learn to identify the common vs. rare versions of the 8 different event types in the videos (top vs bottom row in Fig. 1A); this was fixed across participants and thus constant across conditions. At the start, they were shown a static image depicting each of the 16 events and how they paired into rare/common alternates, but not which were which. Following a preview phase (the first 75 events of the first video of the first object, about 1 min), they were asked to press 'o' if the event was common and 'r' if it was rare, as soon as the event began.

To ensure that participants had understood what was meant by an individual 'event' in the continuous stream, and also that the browser was able to register their key responses, they performed a response-practice task following the preview, in which they pressed the space bar every time a new event started. They were shown a random subsequence of 20 events from the first video and could only move forward once they achieved at least 75% accuracy. Failure on this could be due either to miscomprehension or to technical glitches in registering responses within the time window of the trial; either was grounds for not proceeding. Participants not passing this criterion after ten attempts were compensated but not allowed to proceed to the rest of the task.

Participants then performed the rare vs. common identification task. Each Sequence was shown as a set of consecutive videos (three videos for Training Sequence A, three videos for Training Sequence B, and two videos for the Test Sequence). Each video took about 4.5 min to play. After each video, overall accuracy and percent of trials responded to was displayed, with a reminder that low accuracy could be due to a low overall response rate. The videos for Training Sequences A and B were first (in counterbalanced order across conditions and event assignments), followed by the Test Sequence, though these were not labeled differently for the participants. A new preview was shown prior to the start of each new Sequence.

After completing the set of videos for a given Sequence, participants were given a forced-choice familiarity test to assess their learning of that Sequence. On each trial, two videos were played consecutively side by side, which each showed a mini-sequence of two events. Participants were instructed to choose which video was more typical or familiar by selecting a button below each one; another button allowed them to replay the two videos in that trial. They had to make a selection to continue; no feedback was given. The questions of interest always presented one strong (high transition probability) pair and one weak (low transition probability) pair.

There were three types of forced-choice questions. The 'critical' questions asked participants to compare the strongly predictive, typical two-event sequences (i.e., a cause followed by its effect, for example tilt followed by light for Sequence A) to event pairs that were atypical (~14% transition probability), but formed strongly predictive pairs during the *other* sequence. In this example, we would show hand tap followed by snow, events which were a cause-effect pair in Sequence B. The central object was always shown in the test items, to cue the right sequence; and tests were presented immediately after sequence exposure. For each sequence, there were four critical questions (two for each cause-effect pair). We expected that our effect would hold on these critical questions, based on our findings from the pilot experiment (see Supplemental Material).

Other questions were shown in order to avoid cuing participants to the actual strong pairs, by balancing the number of times the weak pairs (i.e., incorrect options) were shown, and to maintain methods consistency with the pilot experiment. These included four questions which compared the strong pairs to pairs which were always weak (e.g., the cause followed by a different ambient event). Two questions asked participants to compare two strong pairs to each other (e.g. R1 to R2 for Sequence A); these questions did not have a correct answer. Finally, 16 questions presented the weak pairs from the critical questions compared to each other, simply to balance the number of times the weak pairs were presented with the strong pairs. Thus, filler questions ensured that correct videos for the hard questions were not presented more often than the incorrect videos, and so that the same event pairs were tested across all sequences. This created 32 total questions.

Although the Training Sequence videos used in the two conditions were identical, it was important to ensure that the generated transition matrices did not, by chance, differ between the Consistent and Inconsistent Test Sequences in ways that would make the critical questions inherently easier or more difficult for one than the other. The transition probability of the strong pairs, minus the transition probability of the weak pairs, for the critical questions were highly similar between the Consistent and Inconsistent Test Sequences (Consistent $M = 0.842$; Inconsistent, $M = 0.845$).

It should be noted that following the first test, participants could anticipate that such tests would appear during the experiment and this could have motivated them to look for cause-effect relations. However, there was no task-based incentive to track how pairings co-varied across sequences.

We additionally measured verbalizable access to what participants learned. These measures were included because it would seem even more convincing that relational sets are learned spontaneously and affect future learning inadvertently if participants cannot overtly describe the structure they learned and thus would be less likely to strategize about their responses to forced-choice tests in such fashion.

After completing the cover task and forced-choice test for both Training Sequences, participants were asked to "describe anything you learned about each of the two objects you saw," in a text box (freeform response question). Following the Test Sequence, they were additionally asked the following freeform response questions: (1) "During the videos (not the questions), did you notice any patterns in the order of events? Did any events seem to follow each other more than randomly, for any of the objects?"; (2) "Did the videos about each of the objects differ from each other, in terms of which events occurred and in what order?"; (3) "Did you notice any similarities or differences between the first two videos and the last one?" Participants were also asked to note any technical glitches they encountered, and enter their demographic information.

2.6. Scoring and attention measures

Scoring of the freeform responses was done as follows: One score tabulated how many of the four predictive relations the participant had correctly described in their responses (0–4). The second score indicated whether participants were aware of the overall structure of the relations among the sequences (0 or 1). Participants were given a score of 1 if they explicitly noted that one or more relations applied to some sequences but not all of them, or, for those in the consistent condition, if they mentioned that Sequence A and the Test Sequence were more similar to each other than they were to Sequence B.

Performance on the cover task (common vs. rare decision) was used as a measure of attention. Participants with lower than a 60% overall accuracy on the task were excluded from the analysis as described in Participants.

Performance on the cover task also determined the participant's performance bonus. For each of the 8 individual videos, accuracy of 75% or above was awarded $0.50, and catching 25% of the rare events across the experiment was awarded another $1.00, for a maximum of $5.00.

All statistics reported are two-tailed, planned comparisons, unless otherwise indicated, with an alpha level of 0.05. Effect sizes are reported for hypothesis-relevant analyses.

3. Results

Participants in the two conditions had comparable performance on the cover task in terms of overall accuracy (Consistent group: $M = 84.5\%$, CI [83.46, 85.95]; Inconsistent group: $M = 85.1\%$, CI [83.75, 86.29]; $p = .558$), average hit rate (Consistent group: $M = 40.5\%$, CI [37.19, 43.85]; Inconsistent group: $M = 41.3\%$, CI [38.13, 44.61]; $p = .753$), and false alarm rate (Consistent group: $M = 6.3\%$, CI [5.25, 7.17]; Inconsistent group: $M = 5.53\%$, CI [4.69, 6.61]; $p = .247$).

Learning of the individual predictive relations for each sequence was assessed with a force-choice test, presented in between blocks of the cover task. On the questions of interest, participants had to choose between sequence snippets depicting a strong (highly likely) pair of events for that sequence, and a pair that was weak (unlikely) for that sequence but strong for others. Participants in both groups showed above-chance accuracy on this test for each sequence (Sequence A: Consistent group, $M = 60.37\%$, SE = 2.21, CI [0.56, 0.65], $t(187) = 4.70$, $p < .001$; Inconsistent group, $M = 60.24\%$, SE = 2.19, CI [0.56, 0.65], $t(187) = 4.68$, $p < .001$; Sequence B: Consistent group, $M = 55.85\%$, SE = 2.05, CI [0.52, 0.60], $t(187) = 2.86$, $p = .005$; Inconsistent group: $M = 59.84\%$, SE = 2.19, CI [0.56, 0.64], $t(187) = 4.48$, $p < .001$; Test Sequence, Consistent group: $M = 60.77\%$, SE = 2.25, CI [0.56, 0.65], $t(187) = 4.79$, $p < .001$; Inconsistent group: $M = 55.45\%$, SE = 1.64, CI [0.52, 0.59], $t(187) = 3.32$, $p = .001$).

This knowledge of predictive relations was largely not verbalizable. When asked to describe any predictive patterns they noticed, participants correctly identified an average of 0.75 relations out of a possible

4. Additionally, only 8 of 376 participants (2%) correctly described the structure of how the relations differed between sequences. This is unlikely due to any unwillingness of participants to reveal this knowledge. An in-lab pilot sample (Supplemental Materials) exhibited a similar lack of verbalizable access, when probed with active verbal debriefing following the experiment. This suggests that participants were unlikely to be responding to forced-choice tests on the basis of a deliberative strategy making explicit use of knowledge of how the relations covaried.

The analysis of interest was whether condition—the consistency of the Test Sequence with the Training Sequence relational structure—-affected forced-choice test accuracy on individual relation knowledge for the Test Sequence, over and above any differences between groups in the Training Sequences. We thus tested whether the Consistent group was more accurate than the Inconsistent group on the Test Sequence, relative to their difference in performance on the Training Sequences. Training Sequences A and B were collapsed to reflect overall training accuracy. A two-way ANOVA with the factors Condition (inconsistent, consistent) and Sequence Type (training, test) revealed no effect of Sequence Type ($F(1, 374) = 0.38$, $MSE = 0.02$, $p = .539$) or Condition ($F(1, 374) = 0.82$, MSE $= 0.05$, $p = .367$), but a significant interaction ($F(1, 374) = 5.33$, $MSE = 0.25$, $p = .022$, partial $\eta^2 = 0.014$). These results are shown in Fig. 4. The significance of the interaction effect was confirmed with a permutation test for ANOVA, $p = .003$. The simple effect of Condition on Test Sequence was marginally significant (t (374) $= 1.91$, $p = .057$, $d = 0.20$; permutation test $p = .087$). This test is less appropriate than the interaction, however, because it does not take into account individual differences in learning ability (which varies widely). Although statistically robust, it must be noted that the effect size of the interaction was small; a Bayesian analysis of the interaction yielded a Bayes factor of 1.426, indicating positive but not strong evidence. As noted in Methods, our design necessitated that each participant had a very brief exposure to the complex learning manipulation; real-life experience can be more substantial. Our confidence in the statistically reliability of the effect is increased by the a priori design based on pilot data.

Post-hoc t-tests were used to probe the nature of the interaction. In the Consistent group, these revealed no significant difference between the Training Sequence score ($M = 57.85\%$, CI [55.10, 60.59]) and the Test Sequence score (60.51%, CI [56.09, 64.92]), $t(187) = -1.10$, $p = .273$, permutation test $p = .311$. In contrast, in the Inconsistent group, the Training Sequence score ($M = 60.31\%$, CI [57.33, 63.29]) was significantly higher than the Test Sequence score ($M = 55.72\%$, CI [52.44, 58.10]), $t(187) = 2.30$, $p = .023$, $d = 0.168$; permutation test $p = .0418$. Thus, the interaction indicated a decline in performance in the Inconsistent group, but no reliable change in the Consistent group.

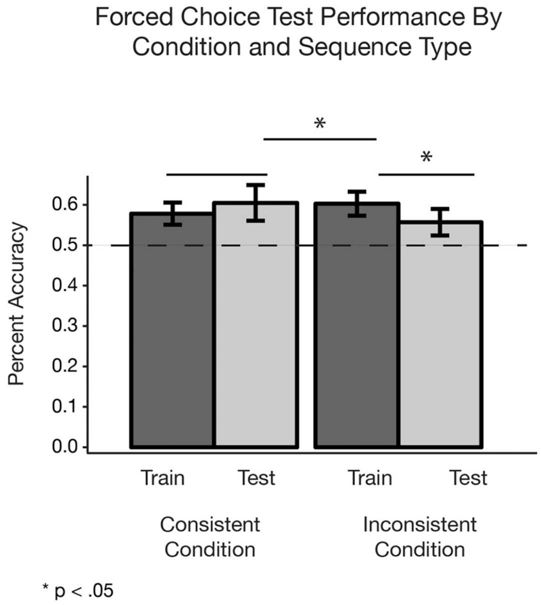

Fig. 4. Results, showing percent accuracy on the critical questions in the forced-choice test by Condition and Sequence type. Error bars indicate 95% confidence intervals using the t distribution. Dotted line depicts accuracy expected by chance (50%). The interaction demonstrates that while Inconsistent group participants performed worse on the Test Sequence than their Training Sequence baseline, the Consistent group did not show such a difference.

Additional tests were performed to rule out alternative explanations of our results. First, we confirmed that there were no differences between the two Training Sequences with an 2 Sequence Type (Sequence A, Sequence B) by 2 Condition (consistent, inconsistent) ANOVA, which revealed no effect of Condition ($F(1, 374) = 0.88$, $MSE = 0.07$, $p = .349$) or Sequence type ($F(1, 374) = 1.19$, $MSE = 0.11$, $p = .277$), nor any interaction ($F(1, 374) = 0.83$, $MSE = 0.08$, $p = .362$), and was confirmed with permutation tests (p's > 0.20). Thus, it was not the case that the Consistent group had, by chance, better performance on the Training Sequence which matched their Test Sequence.

Second, we wished to rule out that the groups differed in their Training Sequence accuracy on the specific predictive relations that differed in their respective Test Sequences. As shown in Fig. 3, the Consistent Test Sequence was shown with R2, while the Inconsistent Test Sequence was shown with R3. Even though the specific events assigned to these relations were perfectly counterbalanced, it is possible that, purely by chance, the groups differed on their knowledge of these particular relations at training. We found this was not the case: there was no reliable difference between the Consistent group's performance on R2 at training ($M = 64.01\%$, CI [58.48, 69.72]) and the Inconsistent group's performance on R3 at training ($M = 60.37\%$, CI [54.68, 66.06]), $t(187) = 0.92$, $p = .359$; permutation test $p = .194$. Thus, the effects of condition cannot be explained by differences in knowledge of the specific individual predictive relations during training.

4. Discussion

Our theories about the world contain not just individual predictive relations, but also knowledge about which relations hang together. Here we investigated whether relations among relations are spontaneously acquired during exposure to events, and whether this knowledge influences later learning. Indeed, we found that observers register how individual predictive relations cohere into higher-order, context-dependent sets, and that this knowledge guides their expectation that these relations will continue to cohere this way in the future—affecting how they process evidence to the contrary. This suggests that binding relations to other, co-varying relations is an intrinsic manner in which the mind encodes the world.

We ensured that learning was over relations, rather than the events composing them, by presenting the same events in each context, but varying which were part of predictive relations (Fig. 3). This follows the classical definition of relational representations, in which relations vary independently of the elements (Gentner, 1983; Markman & Gentner, 1993). We describe this learning as spontaneous, because there was no task demand to learn pairings among the relations. Furthermore, these higher order representations of relational sets inadvertently affected later learning accuracy for individual relations: having learned how pairs of relations co-apply across two contexts, learning was worse when a third context violated this pairing than when it upheld it. This extends prior work on explicit reasoning about how relations cohere (Gershman, 2017; Schulz et al., 2008; Waldmann et al., 2010) to show that it is an automatically operating part of our how we learn even in absence of deliberative reasoning. It is also in line with work in other domains, such as reinforcement learning, showing that correlational structure is inferred even when costly and unnecessary (Collins, 2017; Collins & Frank, 2016).

Our findings have relevance for theories of learning more broadly.

Predictive learning is the backbone of associative learning (Shanks, 1995); and human and non-human animals alike infer the context-sensitivity of relations—for example, that a tone can predict a shock in one room but not another. In certain circumstances, it would appear that contexts serve as 'occasion setters' (Bouton & Swartzentruber, 1986; Urcelay & Miller, 2014): animals learn not that a specific room predicts the shock, but rather, that it modulates the tone-shock relation. In these circumstances, it could be assumed that the room is a cue toward the relation; analogously, in our task, objects can be seen as context cues in just this way. Indeed, what cue serves as a 'context' may well be any stimulus at the top of a predictive hierarchy that predicts more local variation among events, and this role could be statistically inferred (Collins & Frank, 2016; Gershman, 2017). Importantly, participants need no overt cue for context: the relations which hold in it are themselves cues. For example, in our task's Test Sequence, no object cues were available; participants had to use one relation to anticipate the other.

Analogous to this situation is the acquisition of task sets in the reinforcement learning literature. When observers learn multiple stimulus-reward contingencies, they track how these contingencies covary. For example, if at the same time A is rewarding while B is punishing, and at other times these fully reverse, monkeys need only observe one relation (e.g., that A is rewarding) to retrieve the other; no overt cue apart from the predictive statistics themselves is necessary (Saez, Rigotti, Ostojic, Fusi, & Salzman, 2015). We show that relation-relation binding also takes place in purely observational learning, and, again, not as a deliberative strategy where such inferences are beneficial, but as a natural outcome of how the mind encodes observed events.

How can we describe the computations the mind performs to accomplish the binding of relations to other relations? At minimum, learners must determine that relations co-vary in systematic ways, and create a latent structure which captures this co-variation. Finally, in a new context, if a similar individual relation is observed to one already attached to a latent variable, its associated relation(s) can be retrieved and anticipated. The first two operations can be described using models of structure learning, in which probabilistic inferences are made about how relations co-vary and how many clusters of co-varying relations there might be (Collins & Frank, 2016; Gershman & Niv, 2012; Kemp et al., 2010). However, recognizing when a new relation is 'similar' to previously learned ones can be more or less trivial. In our case, recognizing a tilt-light relation in the context of a new object is possible by the visual similarity of these events. But sometimes relations hold in a way that conflicts with visual similarity. In those cases, it is possible that a process of analogical mapping is required to enable inference (Falkenhainer, Forbus, & Gentner, 1989; Gentner, 1983).

A prior step to these is also important. To either map or cluster relations, there must be explicit representations of relations. In our task, it is not enough to keep track the covariation among visible events, since all events occur equally often in all contexts; learners must track the covariation among relations per se. One must therefore suppose a mechanism which creates new latent variables (in a Bayesian framework) or hidden nodes (in a connectionist network framework) which represent the relations themselves (i.e., a variable that represent the correlation between A & B, separately from the stimuli identities). The literature on acquired equivalence suggests such hidden nodes are a natural outcome of predictive learning using a multi-layer autoencoder network (Gluck & Myers, 1993; Honey, Close, & Lin, 2010).

Thus, at this general level of description, our data support the possible existence of a mechanism that forms latent variables to represent relations, makes unsupervised inferences regarding how those relations co-vary, and supports the recognition of similar relations in new contexts in order to retrieve those associates. In future work, we hope to adjudicate between specific alternative implementations of such mechanisms. Moreover, the observation that learning was incidental and not easily verbalizable raises the question of what reliance it may or may not have on working memory or other executive resources, another important topic for future research. Finally, although the relations we studied here were predictive, which are critical and pervasive across many learning tasks, it is possible that different phenomena exist among other relations, such as those representing relative features (*brighter than*, *larger than*; Corral & Jones, 2014) or spatial properties (*below*, *above*). If similar inferential architectures underlie inference of predictive and other relations, then similar principles may apply. Indeed, our work extends the finding of Corral and Jones (2014) that pairs of relations among items are better learned when one of the items is involved in both relations (e.g., A & B and B & C) than when items are not shared. We show here that relations that co-vary consistently, without a shared item, also have an advantage.

More broadly, our claim is that this form of learning is relevant to theory-building. Do the resulting representations have the character of relations composing theories? Have our learners now acquired a 'theory' of two object kinds, composed of their two relations?

One important property of theories is that their descriptions of the world are in a different "vocabulary" than the evidence (Gopnik & Meltzoff, 1997). In our theory of mind, human behavior is not represented as reaches of arms and direction of gaze, but as thought, desire, and belief. The computational work on structure learning cited above offers one formalization of what this means: that theoretical terms are latent variables postulated to explain the evidence regarding how clusters of events or properties cohere (Collins & Frank, 2013; Gershman, 2017; Gershman & Niv, 2012). A latent variable does not refer to an observable event, but rather, to a relation among observables: it captures the fact of their co-variation. One possibility is that elements in theories are exactly such latent constructs, which is why they seem to be in a "different vocabulary": a *belief* explains the coherent co-variation between certain classes of actions. Under this account, our participants created novel latent variables for each of the training sequences, which captured the principle that their two relations hung together. This latent variable was responsible for the expectation that they would hang together later, and thus served as an element in the theory.

Carey (2009) argues that representations inaccessible to awareness, like those we describe, are not conceptual, nor theories. It is thus possible that what we describe are only proto-theories until they are brought into awareness. We nonetheless argue that these representations are highly useful for theory building, particularly for intuitive theories that seem to be formed without substantial deliberative reasoning.

This is not, in any way, a deflationary account of theorizing. It is instead an inflationary account of incidental learning, in line with demonstrations of its ability to generate structured representations of sorts useful for learning linguistic syntax or morphology (Endress, Cahill, Block, Watumull, & Hauser, 2009; Fitch & Hauser, 2004; Friederici, Bahlmann, Heim, Schubotz, & Anwander, 2006; Gerken, 2006; Gomez, Gerken, & Schvaneveldt, 2000; Kovács & Mehler, 2009; Marcus, Vijayan, Rao, & Vishton, 1999; Morgan & Newport, 1981). There are both parallels and differences between the kind of learning described here and the mechanisms supporting grammar learning in natural or artificial languages. Experiments in artificial grammar as cited above have demonstrated the ability of infants and adults to learn complex relations such as the difference between an AAB pattern (first two elements repeat) vs. ABA (the first and last elements match); and A^nB^n patterns, where the same number of A as B elements to follow each other (v.s. A^nB^m patterns). Such relations require a 'phrase structure' grammar, more complex than what is needed to learn relations based on transition probabilities (as here), which can be explained with a simpler, finite state grammar (Fitch & Friederici, 2012; Fitch & Hauser, 2004; Hauser, Chomsky, & Fitch, 2002). Our 'grammar', however, was complex in another important sense: it required representing the co-occurrence of different relations over the same events across contexts, not just the individual relations. This could be captured by two finite state grammars, each specifying two different relations over the same events; these grammars might then be selected based on context (e.g., Gebhart, Aslin, & Newport, 2009; Kovács & Mehler, 2009). The relation between learning grammars required by natural languages, and theories relating predictive structures among events, is a fascinating direction for future research.

5. Conclusion

How theories are learned is a major challenge for cognitive science (Gerstenberg & Tenenbaum, 2017; Tenenbaum et al., 2011). We tackle just one facet of theories: the coherence among multiple predictive relations. We demonstrate that human learners have an inclination to encode higher order relations—how pairs of individual relations themselves cohere—even when these are incidental to the task, and we show that this forms part of the very process by which the mind learns about the world. We argue that this inclination may be a mechanism which spontaneously generates novel constructs to explain observations. It thus forms an important part of our cognitive repertoire, and may explain how we so readily generate intuitive theories.

CRediT authorship contribution statement

Anna Leshinskaya:Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing - original draft, Writing - review & editing, Visualization, Supervision.**Mira Bajaj**:Methodology, Software, Formal analysis, Investigation, Writing - original draft, Visualization.**Sharon L. Thompson-Schill**:Conceptualization, Supervision, Funding acquisition, Writing - review & editing.

Acknowledgements

This work was supported by NIH grants R01DC015359 and R01DC015359 to S.L.T-S, and a Velay Fellowship from the University of Pennsylvania to M.B. This work was previously presented at the 2018 Annual Meeting of the Vision Sciences Society.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2020.104238.

References

Baker, C. L., Saxe, R. R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition, 113*(3), 329–349.

Baker, C. L., Saxe, R. R., & Tenenbaum, J. B. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. *Proceedings of the Cognitive Science Society, 33*(33).

Bouton, M. E., & Swartzentruber, D. (1986). Analysis of the associative and occasion-setting properties of contexts participating in a Pavlovian discrimination. *Journal of Experimental Psychology: Animal Behavior Processes, 12*(4), 333–350.

Carey, S. (2009). *Origin of concepts.* Oxford: Oxford University Press.

Collins, A. G. E. (2017). The cost of structure learning. *Journal of Cognitive Neuroscience, 26*(3), 194–198. https://doi.org/10.1162/jocn.

Collins, A. G. E., & Frank, M. J. (2013). Cognitive control over learning: Creating, clustering, and generalizing task-set structure. *Psychological Review, 120*(1), 190–229.

Collins, A. G. E., & Frank, M. J. (2016). Neural signature of hierarchically structured expectations predicts clustering and transfer of rule sets in reinforcement learning. *Cognition, 152*, 160–169.

Corral, D., & Jones, M. (2014). The effects of relational structure on analogical learning. *Cognition, 132*(3), 280–300.

Dennett, D. C. (1987). *The intentional stance.* Cambridge, MA: MIT Press.

Endress, A. D., Cahill, D., Block, S., Watumull, J., & Hauser, M. D. (2009). Evidence of an evolutionary precursor to human language affixation in a non-human primate. *Biology Letters, 5*(6), 749–751.

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial Intelligence, 41*(1), 1–63.

Fitch, W. T., & Friederici, A. D. (2012). Artificial grammar learning meets formal language theory: An overview. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*(1598), 1933–1955.

Fitch, W. T., & Hauser, M. D. (2004). Computational constraints on syntactic processing in a nonhuman primate. *Science, 303*(2004), 377–380.

Friederici, A. D., Bahlmann, J., Heim, S., Schubotz, R. I., & Anwander, A. (2006). The brain differentiates human and non-human grammars: Functional localization and structural connectivity. *Proceedings of the National Academy of Sciences, 103*(7), 2458–2463.

Gebhart, A. L., Aslin, R. N., & Newport, E. L. (2009). Changing structures in midstream: Learning along the statistical garden path. *Cognitive Science, 33*(6), 1087–1116.

Gelman, S. A., & Wellman, H. M. (1991). Insides and essences: Early understandings of the non-obvious. *Cognition, 38*, 213–244.

Gentner, D. (1983). Structure mapping: A theoretical framework for analogy. *Cognitive Science, 7*(2), 155–170.

Gerken, L. (2006). Decisions, decisions: Infant language learning when multiple generalizations are possible. *Cognition, 98*(3), B67–B74.

Gershman, S. J. (2017). Context-dependent learning and causal structure. *Psychonomic Bulletin & Review, 24*(2), 1–25.

Gershman, S. J., & Niv, Y. (2012). Exploring a latent cause theory of classical conditioning. *Learning & Behavior, 40*, 255–268.

Gerstenberg, T., & Tenenbaum, J. B. (2017). Intuitive theories. In M. Waldmannn (Ed.). *The Oxford handbook of causal reasoning* (pp. 515–548). Oxford: Oxford University Press.

Gluck, M. A., & Myers, C. E. (1993). Hippocampal mediation of stimulus representation: A computational theory. *Hippocampus, 3*(4), 491–516.

Gomez, R. L., Gerken, L., & Schvaneveldt, R. W. (2000). The basis of transfer in artificial grammar learning. *Memory & Cognition, 28*(2), 253–263.

Gopnik, A. (1996). The scientist as child. *Philosophy of Science, 63*(Dec), 485–515.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories.* Cambridge, MA: MIT Press.

Gopnik, A., & Wellman, H. M. (1994). The theory theory. In L. A. Hirschfeld, & S. a Gelman (Eds.). *Mapping the mind: Domain specificity in cognition and culture* (pp. 257–293). Cambridge, UK: Cambridge University Press.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The Faculty of Language: What is it, who has it, and how did it evolve? Author(s): Marc D. Hauser, Noam Chomsky and W. *Tecumseh Fitch Source: Science, 298*(2002), 1569–1579.

Honey, R. C., Close, J., & Lin, T.-C. E. (2010). Acquired distinctiveness and equivalence: a synthesis. *Attention and Associative Learning: From Brain to. Behaviour,* 159–186.

Keil, F. C., Smith, W. C., Simons, D. J., & Levin, D. T. (1998). Two dogmas of conceptual empiricism: Implications for hybrid models of the structure of knowledge. *Cognition, 65*(2–3), 103–135.

Kemp, C., Tenenbaum, J. B., Niyogi, S., & Griffiths, T. L. (2010). A probabilistic model of theory formation. *Cognition, 114*(2), 165–196.

Kovács, A. M., & Mehler, J. (2009). Flexible learning of multiple speech structures in bilingual infants. *Science, 325*(5940), 611–612.

Lombrozo, T. (2009). Explanation and categorization: How "why?" informs "what?". *Cognition, 110*(2), 248–253.

Marcus, G. F., Vijayan, S., Rao, B., & Vishton, P. (1999). Rule learning by seven-month-old infants. *Science, 283*(January), 77–80.

Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology, 25*(4), 431–467.

Morgan, J. L., & Newport, E. L. (1981). The role of constituent structure in the induction of an artificial language. *Journal of Verbal Learning and Verbal Behavior, 20*(1), 67–85.

Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences, 105*(7), 2745–2750.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *The Behabioral and Brain Sciences, 4*, 515–526.

Reber, A. S. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General, 118, 219–235, 118*(3), 219–235.

Saez, A., Rigotti, M., Ostojic, S., Fusi, S., & Salzman, C. D. (2015). Abstract context representations in primate amygdala and prefrontal cortex. *Neuron, 87*(4), 869–881.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by eight-month-old infants. *Science, 274*(5294), 1926–1928.

Schulz, L. E., Goodman, N. D., Tenenbaum, J. B., & Jenkins, A. (2008). Going beyond the evidence: Abstract laws and preschoolers' responses to anomalous data. *Cognition,* 1–48.

Shanks, D. R. (1995). *The psychology of associative learning.* Cambridge, UK: Cambridge University Press.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science, 331*(6022), 1279–1285.

Turk-Browne, N. B., Jungé, J., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology. General, 134*(4), 552–564.

Urcelay, G. P., & Miller, R. R. (2014). The functions of contexts in associative learning. *Behavioural Processes, 104*, 2–12.

Waldmann, M. R., Meder, B., Von Sydow, M., & Hagmayer, Y. (2010). The tight coupling between category and causal learning. *Cognitive Processing, 11*(2), 143–158.

Werchan, D. M., Collins, A. G. E., Frank, M. J., & Amso, D. (2015). 8-month-old infants spontaneously learn and generalize hierarchical rules. *Psychological Science, 26*(6), 805–815.