Incremental Hypothesis Revision in Causal Reasoning Across Development

Rebekah Gelpi[1] **(rebekah.gelpi@mail.utoronto.ca),**
Ben Prystawski[1] **(ben.prystawski@mail.utoronto.ca),**
Christopher G. Lucas[2] **(clucas2@inf.ed.ac.uk),**
and Daphna Buchsbaum[1] **(buchsbaum@psych.utoronto.ca)**

[1]Department of Psychology, University of Toronto, Canada
[2]School of Informatics, University of Edinburgh, United Kingdom

Abstract

We explore whether children's strategies on a causal learning task show a bias observed in adults towards "exploitative" hypothesis revision. Adults and children (ages 4–6) were presented with evidence which initially seemed to conform to a simple, salient rule (e.g. blue blocks activate a machine), but then encountered evidence that violated this rule. The true rule in the "near" condition was more complex, but could be reached through iterative revision of the salient rule, while in the "distant" condition, the true rule was comparatively simple, but incremental revision could not yield the true rule. Participants then predicted the behaviour of a set of new blocks. Adults performed better in the near condition, while in the distant condition adults did not appear to revise their initial hypothesis significantly. Unlike adults, children's overall performance in both conditions was similar, while condition differences may reflect a broader search for alternative solutions.

Keywords: Bayesian inference, belief revision, causal learning, process model

Introduction

Why does a boat float, but a pebble sink? To young children, weight and density are undifferentiated, failing to explain why some large, heavy objects can nevertheless float; as children develop, they must—and do—dramatically shift their beliefs to fully explain the evidence they encounter in the world (Smith, Carey, & Wiser, 1985). This flexibility is a critical component of human learning and development, but too much willingness to shift one's beliefs could result in abandoning a belief which could nevertheless be correct, or could be incrementally improved. However, insufficient revision of one's views could prompt excessive attachment to a belief even after it has been shown to be wrong.

Despite the complexity and ambiguity of data encountered in the world, humans are nevertheless able to effectively filter incoming information, and generalize this information to make broad conclusions. One approach successfully used to predict rapid human knowledge acquisition in tasks such as causal inference is the use of ideal-observer hierarchical Bayesian models (Kemp, Perfors, & Tenenbaum, 2007; Lucas & Griffiths, 2010; Tenenbaum, Kemp, Griffiths, & Goodman, 2011). Optimal Bayesian learners are able to represent the posterior distribution in its entirety, allowing for the exact computation of the probability of various hypotheses given some observed data by using Bayes' rule. While these models often accurately reflect the group-level, if not individual, performance of humans on certain tasks, the idealized computations quickly become time-consuming and computationally expensive, and are often intractable in real-world tasks as the number of possible hypotheses an ideal Bayesian learner would need to maintain grows with new data, and outstrips humans' limited cognitive capacities. As a result, we might expect human performance to deviate from the Bayesian ideal in complex tasks where computing the full posterior becomes intractable. Nonetheless, these optimal inferences can often be approximated with algorithms that are efficient and accurate most of the time, but which nevertheless show certain systemic biases, such as anchoring and base-rate neglect (Lieder, Griffiths, Huys, & Goodman, 2017; Sanborn & Chater, 2016). Rather than always attempting to determine the optimal choice or inference regardless of the potentially overwhelming cost of doing so, human beings may make *resource-rational* decisions, which are optimal given the bounded nature of human cognition and the costs of making inferences (Lieder & Griffiths, 2020).

One way to represent the algorithmic processes involved in Bayesian inferences is to use methods which generate samples from the posterior distribution rather than maintaining a full posterior over all possible outcomes. In particular, Markov chain Monte Carlo (MCMC) sampling methods (e.g., Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Griffiths, Lieder, & Goodman, 2015; Sanborn, Griffiths, & Navarro, 2010) sequentially generate new samples that incrementally change a current hypothesis in a way that asymptotically approximates the posterior distribution. The incremental nature of these algorithms means that they can weight initial hypotheses more than a normative model would predict, if one draws a small number of samples. Nonetheless, given that sampling can be costly, making apparently sub-optimal inferences based on few, or even just a single sample may be an optimal long-term strategy for some decisions (Vul, Goodman, Griffiths, & Tenenbaum, 2014).

Resource-rational models of cognition are consistent with findings of a bias towards exploiting existing knowledge in adults' hypothesis search strategy (Herbst, Lucas, & Buchsbaum, 2017) and a preference towards local, conservative edits to one's working hypothesis (Bramley, Dayan, Griffiths, & Lagnado, 2017). An incremental approach is helpful for fine-tuning good hypotheses, but too much conservatism is maladaptive when one's starting state is fundamentally wrong—as with a child who has no concept of density. If children are more aggressive and less incremental in their belief revision as Gopnik et al. (2017) have speculated, it could explain their ability to outperform adults in some tasks, especially

Transition Rule	$p(t)$
$S \rightarrow \forall x, \ell(x) \Leftrightarrow (D)$	1
$D \rightarrow (C) \vee D$	τ
$D \rightarrow (C)$	$1 - \tau$
$C \rightarrow (P) \wedge C$	τ
$C \rightarrow (P)$	$1 - \tau$
$P \rightarrow FEF$	θ
$P \rightarrow V$	$1 - \theta$
$V \rightarrow F$	0.5
$V \rightarrow \neg F$	0.5
$F \rightarrow \{A, B, C, D, E\}$	β_{A-E}
$E \rightarrow =$	0.5
$E \rightarrow \neq$	0.5

Table 1: Transition rules in the modified DNF rule grammar. $\ell(x)$ returns true if x is a category member and false otherwise. D denotes a disjunction of conjunctive clauses. C denotes a conjunction of propositions which are treated as literals in the grammar. E refers to an equality or inequality ($=$ or $\neq$) and F refers to a block feature. τ is the transition probability. θ is the probability that a transition will introduce an equality or inequality. Each feature has an associated β which represents the probability that the feature is relevant to a given rule.

ones that require choosing a counter-intuitive and unfamiliar solution over a familiar one with less explanatory power (Lucas, Bridgers, Griffiths, & Gopnik, 2014; Gopnik, Griffiths, & Lucas, 2015). The developmental change observed in search strategies on these tasks may resemble the process of simulated annealing, wherein searches begin at high "temperatures" (variability) and gradually "cool" as they narrow in scope to the most likely possibilities (Gopnik et al., 2015).

Under this view, apparently noisy or irrational beliefs in children are not simply the result of a failure to recognize that a hypothesis is reliable, nor of differing prior beliefs, but an adaptive policy for learning, given children's particular cognitive limitations. Just as adults' tendency to exploit information that is already available is often a sensible allocation of limited resources, children may choose to explore, and in doing so learn more about the world, gaining the knowledge that they can, as adults, quickly exploit (Denison, Bonawitz, Gopnik, & Griffiths, 2013).

To explore this idea, we compared adults' and children's performance in a task where participants must integrate counter-intuitive evidence, following evidence that supports a simple, salient causal rule, extending Herbst et al. (2017). We propose that children, unlike adults, will not show better performance in a condition rewarding incremental local hypothesis edits. We use a Bayesian model to illustrate how an ideal observer would revise their beliefs to account for evidence in these tasks. By comparing participants' judgments to the predictions of the model, as well as to alternative deterministic rule-based strategies, we can reveal developmental differences with an eye to more or less incremental belief revision. If children are exploring possible solutions to the task more broadly, their judgments will not simply be a noisy reflection of those made by adults, but may instead use different strategies that yield qualitatively different judgments.

Bayesian Model of Causal Inference

To understand how an ideal Bayesian learner would make inferences about a causal rule based on observations, we adapted the Bayesian rule induction model from Goodman et al. (2008). This model infers a distribution over rules that define category membership as a function of an object's features, based on example objects that are or are not in the category. The model favours rules that are simpler, as well as those that best explain the data. This distribution over rules can then be used to infer the category of new objects, and the relative probabilities of objects being in the category.

In this model, rules are logical propositions generated using a context free grammar (Table 1). The propositions are expressed in a modified disjunctive normal form (DNF): that is, rules are composed of disjunctions of conjunctions of predicates (P). Our model deviates from standard DNF in that we allow equality or inequality between features at the bottom level of an expression, while traditionally DNF uses only features or their negations in clauses.

The model assigns a prior probability to each rule according to its syntactic complexity, where more complex rules are less probable:

$$P(r) = \prod_{t \in Deriv(r)} p(t) \qquad (1)$$

where $Deriv(r)$ is the set of all transitions in the derivation of rule r and $p(t)$ is the probability associated with transition t.

We generated model predictions for the tasks in our experiments. The model observes a series of ten blocks, each with five binary features, labelled A through E. The feature A represents the background colour of the block, and the other four features represent individual shapes on the block. Each of the blocks does or does not activate a machine, and the model infers the probability that a given rule, based on the features of the blocks, determines the observed pattern of causal activations. Blocks that activate the machine are known as *blickets*. We assume that deterministic rules are most likely, as both adults (Frosch & Johnson-Laird, 2011) and children (Schulz & Sommerville, 2006) strongly prefer deterministic causal relationships. However, there is a small probability, $\varepsilon = 0.001$, that the machine will misfire and a block will be an exception to the rule (e.g., a block activates the machine despite the rule predicting it should not). These assumptions give us the posterior distribution over rules:

$$P(r|\mathbf{B}, \mathbf{L}) \propto \varepsilon^k (1 - \varepsilon)^{n-k} p(r) \qquad (2)$$

where r is a possible causal rule, $\mathbf{B}$ is the set of blocks being observed, $\mathbf{L}$ is the observed pattern of activations for the blocks in $\mathbf{B}$, ε is the noise parameter, n is the total number of observed blocks, and k is the number of outliers whose activations do not match the rule r's prediction.

All of our model predictions are given for values of $\tau = 0.4$ representing a modest simplicity bias towards shorter rules, and and $\theta = 0.2$, a moderate probability that the rule will directly compare two features, but our model predictions are
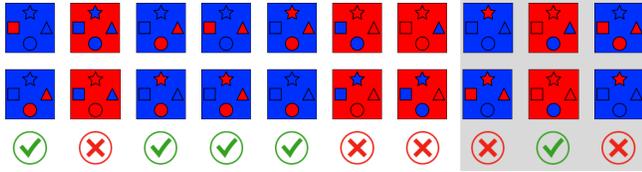
Figure 1: Sample stimuli and block order for Experiment 2's blicket detector task in the distant (top) and near (bottom) conditions. Blocks have the salient background feature (*A*), two rule-relevant features (the circle, *C*, and square, *D*), and two rule-irrelevant features (the star, *B*, and triangle, *E*). The first seven blocks appear to conform to an initially salient, simple rule, e.g. "blue blocks are blickets". However, the final three blocks violate this intuition, revealing that the true rule requires revising one's beliefs to correctly predict all machine activations.

robust across a relatively broad range of parameter values: 0.3 to 0.5 for τ, 0.1 to 0.25 for θ, and 0.001 to 0.05 for ε. Additionally, we assume that the feature space will be sampled roughly in proportion to the size of the feature, favouring the salient feature with a probability of 0.5 (and other features equally at 0.125 each).

In our task, we present ten blocks whose activation patterns are deterministically captured by a single rule: the *true rule*. The first seven blocks are also consistent with a simpler, *a priori* likely rule: the *initial rule*. However, the last three blocks are outliers under this rule, and are only explained by the true rule (Figure 1). There are two conditions for the experiment, in which the true rule governing the pattern of activations differs: the *near* condition and the *distant* condition. In both conditions, the initial rule remains the same; namely, that the background feature of the block *A* determines whether it is a blicket. In the near condition, the true rule is most simply written as $(A \wedge \neg C) \vee (\neg A \wedge \neg D)$. This rule incorporates the initial rule (*A*), and it can be reached through a process of iterative hypothesis revision, by making single edits to a working hypothesis that steadily improve one's accuracy (e.g., editing the rule (*A*) which explains 7/10 examples into $(A \wedge \neg C)$, which explains 9/10). In contrast, in the distant condition, the simplest true rule which could accurately categorize all blocks in the exposure phase can be written $(C \neq D)$. This rule does not involve the background colour, and there are no incrementally better hypotheses that provide a stepping stone to the true hypothesis. Instead, it is necessary to temporarily adopt a less useful rule (e.g., $A \wedge C$, which explains 2/10 blocks) in order to reach the true rule via single edits, or to dismiss one's initial hypothesis about the background colour as the most likely rule entirely.

Model Predictions and Discussion

We estimated the posterior distribution that an ideal Bayesian observer would infer over rules after each of the 10 blocks, using a Metropolis-Hastings algorithm (ten chains of 60,000 samples; burn-in = 10,000). In both conditions, the initial rule is the most probable rule until introduction of the inconsistent evidence, at which point the initial rule rapidly becomes unlikely, and the true rule becomes the most probable, even when it was highly unlikely *a priori* (Figure 2). In both con-

ditions, the model assigns the highest posterior probability to the true rule by the time all 10 blocks are observed. However, in the distant condition, the true rule becomes the only plausible rule, while in the near condition the model remains less certain, as the intermediate rule $A \wedge \neg C$ (a necessary component of the true rule), and a simpler rule, $(\neg C)$, that requires dismissing *A* can each account for 9 of the 10 observations. In the following sections we present both adult participants and 4-6 year old children with this same task.

Experiment 1: Adult Hypothesis Revision

In Experiment 1 we present a causal inference task to test whether adults' revision process is exploitation-biased. If adults adjust their beliefs in a way that is consistent with our ideal observer, they should be at least as likely to dismiss the initial hypothesis that the background determines block activations in the distant condition as in the near condition. However, if adults' hypothesis exploration is conservative, favouring local edits to a working hypothesis that steadily improve accuracy (and potentially reducing the search space for revisions to their beliefs), then adults will not change their beliefs in the distant condition, where the edits required to reach the true rule force one to abandon a local optimum. Thus, we predict that adults' judgments will reflect this process of revision in the near but not the distant condition, where they will fail to develop a better rule than the initially induced hypothesis.

Methods

Participants and Design   118 adult US residents were recruited through Amazon's Mechanical Turk service and paid \$0.50 for completing the task. 28 participants were excluded due to failure to correctly answer one or more attention checks. 90 participants were included in the final analysis, divided between near ($N = 45$) and distant ($N = 45$) conditions. Sample size was determined using Frick's (1998) sequential analysis criteria, with a lower bound of 90 participants.[1]

Materials and Procedure   Adults completed an online causal learning task. They first watched a short video that introduced them to a machine called a blicket detector. When blocks known as blickets are placed on the device it plays a sound, displays a check mark, and animates. When blocks that are not a blicket are placed on the machine, it remains inactive. As in Herbst et al. (2017), blocks had five features: the background colour *A*, two rule-relevant features *C* and *D*, and two irrelevant features *B* and *E*.

Participants were then presented with an exposure phase in which they dragged 10 blocks one at a time onto the detector, which would either activate or not. The block was then added to a visible list of blickets or non-blickets. As with our model, the first 7 blocks appeared to follow the initial rule, while the last 3 blocks violated it, and the true rule for the condition (near or distant) accounted for these seeming exceptions. All participants within a condition saw the same 10 blocks (5 blickets and 5 non-blickets), with a randomized

---

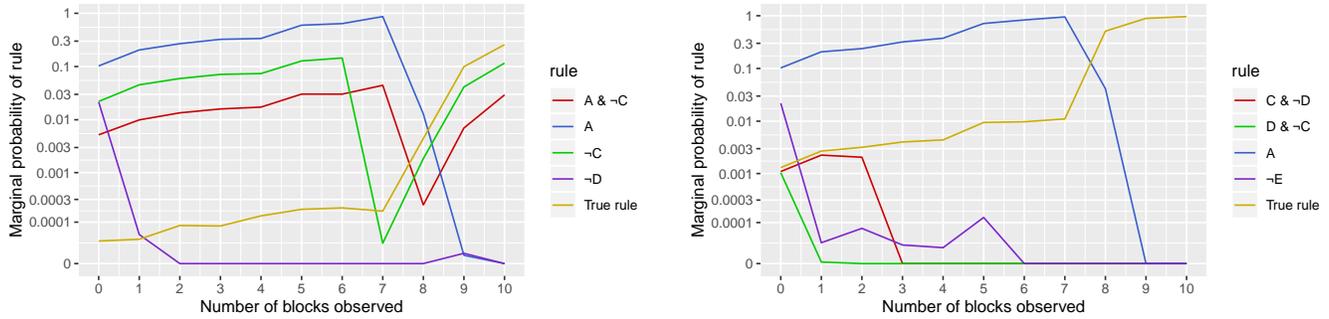[1]Our analyses were preregistered and can be found here: osf.io

Figure 2: Rule probabilities by number of observations. In the near condition (left), the true rule was most probable under the model, although other rules which explained 9/10 blocks were also assigned non-negligible probabilities. In the distant condition (right), the true rule's final probability was near 1.

order within the first 7 (initial rule-consistent) and final 3 (initial rule-violating) blocks. The role of blue and red (e.g., as background colour according to the initial rule) was counterbalanced.

Adult participants then completed two tasks: a blicket identification task and a forced-choice task. In the blicket identification task, participants were presented with 8 blocks and asked to judge whether the block was a blicket, and to indicate their confidence in this judgment on a seven-point scale. The blocks presented consisted of four blickets and four non-blickets according to the true rule. Two of each were consistent with the initial rule and two violated the initial rule.

The blicket identification task was followed by a forced-choice task. In the forced-choice task, participants were presented with two blocks on each trial, a blicket and a non-blicket, and asked to select which block they thought was the blicket. Both the blicket and the non-blicket could either be initial rule-consistent (e.g., for initial rule A = blue, a blue background blicket and a red background non-blicket), or initial rule-violating (a red background blicket and a blue background non-blicket). Accounting for all possible combinations of blicket and non-blicket initial rule consistency, this yielded a total of four trials for the forced-choice task. Finally, adults were asked a series of attention and comprehension checks, to ensure that they had attended to the entire experiment.

Results and Discussion

The results supported our prediction that adults would score more highly in the near condition than the distant condition, consistent with an incremental, exploitation-biased hypothesis search strategy (Figure 3). Participants in the near condition performed significantly better than those in the distant condition (z = –2.81, χ^2 = 8.08, p = 0.005). Across conditions, participants were better at categorizing both blickets (z = 4.57, χ^2 = 22.21, p < 0.001) and non-blickets (z = 5.09, χ^2 = 27.78, p < 0.001) correctly when the blocks were consistent with the initial rule. A 2-way interaction was found between non-blicket consistency and condition (z = 2.53, χ^2 = 6.52, p = 0.011); participants performed significantly worse in the distant condition than in the near condition when the non-blicket looked like a blicket according to the initial rule.

No other interactions were significant (all p > 0.27).

To investigate the interaction, we ran follow-up pairwise comparisons by condition. When the non-blicket was consistent with the initial rule (e.g. for initial rule A = blue, a red background non-blicket), there was no condition difference (z = 0.20, p = 0.84); however, participants were more accurate in the near condition than in the distant condition when the non-blicket violated the initial rule (z = 3.74, p < 0.001). If participants use the initial rule to categorize blocks, they would consider initial rule-violating non-blickets to be blickets, and select them as often as the true blicket when the true blicket is consistent with the initial rule, and more often than the true blicket when both blocks violated the initial rule. This appears to describe performance in the distant condition (Figure 3). Participants do not appear to do this to the same degree in the near condition, suggesting that some participants have learned the true rule, or an intermediate rule: for example, that blocks must additionally satisfy the constraint ($A \land \neg C$) in order to be blickets.

For the blicket identification task, a 2x2x2 mixed ANOVA was run, with initial rule consistency, blicketness, and condition as factors. As in the forced-choice task, participants were more accurate at correctly identifying blocks in the near condition than in the distant condition, $F(1,88)$ = 5.05, p = 0.027. Participants were no better at categorizing blickets than non-blickets, $F(1,88)$ = 0.80, p = 0.375. In both conditions, they also more accurately categorized initial rule-consistent blocks than initial rule-violating blocks, $F(1,88)$ = 58.3, p < 0.001. There were no significant two-way interactions (all p > 0.09), but the analysis found a three-way interaction, $F(1,88)$ = 8.35, p = 0.005. As was the case with the forced-choice task, participants in the near condition were better at recognizing initial rule-violating non-blickets (e.g. blue background non-blickets) than in the distant condition, $t(88)$ = 2.34, p = 0.022. Participants in the near condition may have been able to reach an intermediate rule that allowed them to correctly categorize initial rule-violating non-blickets, even though they were no better at categorizing initial rule-violating blickets (e.g. red background blickets), $t(88)$ = 0.28, p = 0.78, in the near condition compared to the distant condition.

We also tested whether adults' responses on the forced-

choice task were best captured by the idealized model's predictions of the probability that each block was a blicket, after marginalizing over possible true rules, or by individual rule predictions (the expected choices if adults were deterministically applying a single specific rule, with small outlier probability ε), by comparing the log likelihoods of adults' judgments under each model. In the near condition, the full posterior model ($\ell = -373.7$) was the best fit; this was followed by the true rule ($\ell = -912.1$), the intermediate rule $A \wedge \neg C$ ($\ell = -953.5$), and finally the initial rule ($\ell = -1036.4$). However, in the distant condition, the initial rule was the closest fit for the adults' judgments ($\ell = -843.0$), followed by the model ($\ell = -958.0$) and the true rule ($\ell = -1257.4$).

Overall, these results suggest that while adults in the near condition are able to revise their beliefs to explain at least some seeming exceptions in the initial rule, in the distant condition they cannot successfully revise their beliefs to generate a hypothesis better than the initial rule. On the other hand, young children's flexibility and openness to unusual solutions may predispose them to explore alternative solutions more broadly (e.g., Lucas et al., 2014; Gopnik et al., 2015, 2017), potentially allowing them to escape the local optimum that adults were unable to in the distant condition.

Experiment 2: Children's Hypothesis Revision

In Experiment 2, we adapted the blicket detector task administered to adults to one that would be suitable for young children, and predicted that the condition differences observed in adults would not be present in children from ages 4-6. As our model indicated that the distant condition rule was more easily reached if one's working hypothesis did not constrain one's search space, we predicted that if a condition difference did emerge, it would likely be reversed from that observed in adults, especially in younger children. Additionally, if children are more willing to abandon the initial rule, their tendency to choose the initial rule-consistent blocks should be weaker than that of adults.

Methods

Participants and Design Data was collected for 99 4-6 year old children ($M = 5.35$, $SD = 0.99$) through in-lab participation and at local museums. Nine children were excluded from the final analysis for failing a comprehension check ($N = 7$), due to family interference ($N = 1$) or experimenter error ($N = 1$), for a total of 90 included in our initial analysis. As with adults, participant sample size was determined through Frick's (1998) criteria, with a lower bound of 90 participants.[2]

Materials and Procedure Children were presented a physical task with eighteen wooden blocks and a blicket detector with an internal chime. The block stimuli used with children were similar to those used with adults, adapted for ease of recognition of individual features by children, and featured a blue or red background and four familiar shapes.

Children were introduced to the blicket detector as a "special machine" which would make a sound if a blicket was placed on it. Like the adults, they observed a series of blocks with 7 blocks consistent with the initial rule followed by 3 blocks that violated the initial rule. As they tested blocks, children were instructed to place blickets to one side and non-blickets to the other; all blocks remained visible until the end of the exposure phase. The conditions used in this experiment were logically equivalent to Experiment 1's, and featured the same 10 exposure blocks. After completing the exposure task, as a comprehension check, children were asked if a blicket would make a sound when put on the machine. Those who failed to answer correctly were excluded from the final analysis.

Children then completed a forced-choice task, which used a predetermined list of eight blocks. Children saw the same 4 trials as adults, with the pairing of blickets and non-blickets presented for each trial was randomized. Finally, children were asked to explain why they chose blocks that they did; if they did not respond, children were prompted to explain why they thought those blocks were blickets. During this time, all blocks from the forced-choice task were visible to the child.

Results and Discussion

The criterion for sequential analysis used to determine sample size for children was the two-way interaction between age (child or adult) and condition on the forced-choice task on a 2x2 mixed-effects logistic regression. There was a significant main effect of age ($z = 2.11$, $\chi^2 = 4.56$, $p = 0.034$), but not of condition ($z = -1.52$, $\chi^2 = 2.37$, $p = 0.13$). The interaction of age by condition did not reach our pre-specified criterion of $p > 0.36$ or $p < 0.01$ ($z = -2.26$, $\chi^2 = 5.22$, $p = 0.023$). Thus, data collection is ongoing, but we report preliminary findings from the first 90 children here.

Unlike adults, children did not differ in their overall performance between the near and distant conditions ($z = 0.69$, $\chi^2 = 0.47$, $p = 0.49$). Across both conditions, they were significantly better at recognizing initial rule-consistent blickets ($z = 3.55$, $\chi^2 = 12.92$, $p < 0.001$) as well as initial rule-consistent non-blickets ($z = 3.14$, $\chi^2 = 10.07$, $p = 0.002$). There were no two-way interactions in the data, but there was a marginal 3-way interaction ($z = 1.90$, $\chi^2 = 3.66$, $p = 0.056$). In contrast to adults, for whom this pattern was reversed, this appeared to be driven by children in the near condition being worse than in the distant condition at choosing the true blicket when both blocks violated the initial rule.

We also examined children's performance relative to chance. In the distant condition, children were above chance at selecting the correct block as the blicket when both blocks were initial rule-consistent, $B(31,45)$, $p = 0.016$; in the near condition this effect was not significant, but in the same direction $B(28,45)$, $p = 0.135$. Further, children in the near condition were significantly below chance when both blocks violated the initial rule, $B(10,45)$, $p < 0.001$, suggesting that in both conditions, children often chose the block with a background colour consistent with the initial rule when given the
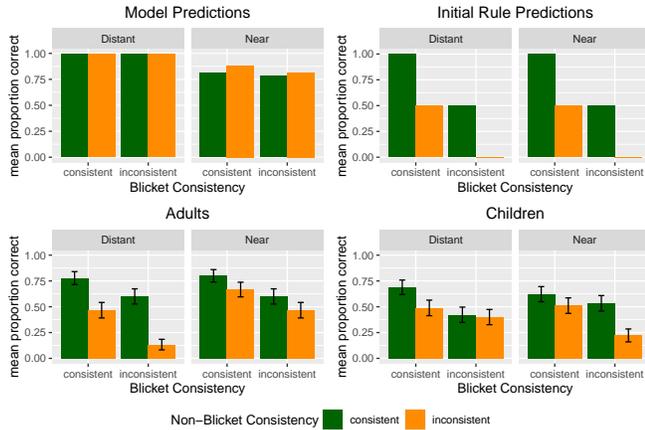
Figure 3: Comparison of model predictions and performance by adults and children. The model's predictions (top left) were the closest match for the true rule in the distant condition, matching the rule with a probability near 1 for all blocks; its predictions remained a good match in the near condition, although it assigned less confidence in these choices. Compared to the model, both adults (bottom left) and children (bottom right) had strong effects of initial rule-consistency; they were less accurate at categorizing initial rule-violating blocks. Unlike children, adults' accuracy was further reduced for initial rule-violating non-blickets (yellow bars) in the distant condition, relative to the near condition. Adults' choices in the distant condition more closely resembled a noisy application of the initial rule (top right), while their choices in the near condition were better fit by the idealized model's predictions.

choice. Interestingly, children in the distant condition were not significantly below chance, $B(18,45)$, $p = 0.233$, when both blocks violated the initial rule.

Together, these preliminary findings support our prediction that children's performance would not be significantly better in the near condition relative to the distant condition, and that their performance would differ from adults by condition. While overall, children were less accurate on the task than adults, they did not merely randomly choose blocks, nor do their choices within each condition resemble those of adults, which may reflect developmental differences in the process by which children revise their beliefs.

As with adults, we tested the likelihood of children's responses against both our idealized model and the application of individual rules. Unlike adults, our model was the likeliest given the data for children in both the near ($\ell = -699.2$) and distant ($\ell = -962.9$) conditions. In the near condition, this was better than the initial rule ($\ell = -1298.8$) and the true rule ($\ell = -1243.6$); likewise, this was true in the distant condition (initial rule $\ell = -1326.5$; true rule ℓ –1209.0). These results are suggestive of a noisier strategy, but importantly, one that differs from the pattern observed in adults.

Finally, we performed follow-up pairwise comparisons to determine whether a developmental difference in strategies emerged within child age groups. 4-year-old children were marginally better in the distant than the near condition ($z = 1.94$, $p = 0.052$), whereas 5- and 6-year-olds were no better in either condition (all $p > 0.28$). Older children may be on the cusp of developing more adult-like strategies in their hypothesis search, becoming less sensitive to unusual evidence

and more reliant on prior beliefs (e.g., Gopnik et al., 2017).

One possibility for at least some of the children's responses is that their hypothesis about which blocks are blickets take the form of a higher-order structure than the features of an individual block as a logical rule. When prompted to explain their choices, a number of children suggested that characteristics such as sequences of blocks with a given background colour or presentation position, such as "red-red-blue-red" or "right-left-right-left" ($N = 5$), determine which blocks are blickets.

While it has been suggested that children may simply begin with different priors about data than adults (Gopnik et al., 2017; Lucas et al., 2014), we believe that children find the initial rule to be similarly salient to adults in our experiment. Suggestively, 30 of 90 children spontaneously articulated a form of the initial rule in the exposure phase (e.g., "oh, it's the blue blocks!"), suggesting that like adults, children provisionally operate according to a form of the initial rule as they observe the initial consistent blocks. This supports the possibility that process-level differences, such as making higher-temperature edits to a working hypothesis than adults, may explain at least some of the changes across development in this task.

General Discussion

These experiments support the idea that adults' hypothesis search strategies are incremental and exploitation-biased, and suggest that children do not exhibit similar exploitative behaviour in generating hypotheses to explain observed data. Consistent with prior research, adults were more accurate in classifying blocks when incremental alterations to an initial salient hypothesis could steadily improve their accuracy. This is despite the fact that an idealized Bayesian model found the true rule in both conditions the most probable, and significantly more probable in the distant condition. Adult participants' better performance in the near condition appeared to be partly driven by greater accuracy at correctly categorizing initial rule-violating non-blickets, consistent with adults incrementally updating their hypotheses to account for the fact that not all blocks that shared a background colour with the initially encountered blickets would activate the machine.

These results are also consistent with the hypothesis that unlike adults, children's search strategies are not clearly exploitation-biased. Children did not perform differently across conditions, but also did not choose randomly: many children quickly and spontaneously understand that the initial blocks they are presented with match a salient and easily observable initial rule, and there is a significant effect of both true blicket and distractor block consistency on their probability of correctly categorizing a block. However, their choices on the task may indicate that their inferences about effective procedures for successful belief revision may be subject to fewer inductive constraints than adults, resulting in more variable solutions that nevertheless might have been no better than the initially induced rule. The presence of U-shaped

learning curves in many domains (Carlucci & Case, 2013) may follow from an increased willingness to abandon a local optimum for what is, in the long term, a better solution.

The probabilistic model we developed identified areas in which an ideal Bayesian observer would systematically diverge from adults. Adults' judgments in the near condition were best fit by this model, while in the distant condition, the initial rule was a likelier predictor of adults' responses, suggesting a bias towards local edits that incrementally improve one's accuracy. This may be sensible when one's existing beliefs are approximately correct, but it can prevent one from being able to escape a "good-enough" strategy if there are no easy edits in reach. While adults may vary in their individual ability to incrementally revise, the good group-level fit of a model estimating the posterior distribution (rather than any individual rule) in the near condition is suggestive of a "wisdom of crowds" effect, consistent with the hypothesis that adults may be sampling from the posterior distribution (Denison et al., 2013; Sanborn & Chater, 2016).

The exact iterative process by which adults are updating their beliefs, and thus how this process fails in the distant condition, remains unclear, as does the method by which children choose to evaluate the evidence they encounter given limited cognitive resources. For example, adults might be falling victim to a "learning trap" (Rich & Gureckis, 2018), disregarding relevant dimensions of a solution. To disambiguate possible processes of belief revision, we plan to administer new versions of the task eliminating the induced order effect, as well as adapt and extend our existing model through the use of sequential Monte Carlo (or particle filter) algorithms to account for the process by which incremental updates are made when participants (and adults in particular) encounter evidence that violates their initial assumptions. Particle filters allow a learner to sequentially reweight and revise beliefs with new data, introducing statistical dependency in one's representation of the posterior distribution (Sanborn et al., 2010). Resource constraints may thus result in few opportunities to improve a working hypothesis that does not provide an opportunity for improvement through local, conservative edits. Higher stochasticity in a child's edit behaviour, on the other hand, may lead to more dead ends as a child considers unproductive changes to a working hypothesis, but might also lead to hypotheses unexplored by adults that could help them find an otherwise unexpected solution more often than adults.

Overall, our findings suggest that adults are better able to correctly learn a rule reachable by a process of iterative revision to a working hypothesis that is at least as complex, if not more so, than a rule which is simple, but requires dismissing one's initial assumptions. Children do not display the same bias in their search strategy, and may be less committed to any one rule as they more broadly explore possible solutions.

Acknowledgements

References

Bramley, N. R., Dayan, P., Griffiths, T. L., & Lagnado, D. A. (2017). Formalizing Neurath's ship: Approximate algorithms for online causal learning. *Psychological Review*, *124*(3), 301–338.

Carlucci, L., & Case, J. (2013). On the necessity of U-shaped learning. *Topics in Cognitive Science*, *5*(1), 56–88.

Denison, S., Bonawitz, E., Gopnik, A., & Griffiths, T. L. (2013). Rational variability in children's causal inferences: The Sampling Hypothesis. *Cognition*, *126*, 285–300.

Frick, R. W. (1998). A better stopping rule for conventional statistical tests. *Behavior Research Methods, Instruments, & Computers*, *30*(4), 690–697.

Frosch, C. A., & Johnson-Laird, P. (2011). Is everyday causation deterministic or probabilistic? *Acta Psychologica*, *137*(3), 280–291.

Goodman, N., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154.

Gopnik, A., Griffiths, T. L., & Lucas, C. G. (2015). When younger learners can be better (or at least more open-minded) than older ones. *Current Directions in Psychological Science*, *24*(2), 87–92.

Gopnik, A., O'Grady, S., Lucas, C. G., Griffiths, T. L., Wente, A., Bridgers, S., ... Dahl, R. E. (2017). Changes in cognitive flexibility and hypothesis search across human life history from childhood to adolescence to adulthood. *Proceedings of the National Academy of Sciences*, *114*(30), 7892–7899.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, *7*(2), 217–229.

Herbst, E., Lucas, C., & Buchsbaum, D. (2017). Investigating the explore-exploit tradeoff in adult causal inferences. *Proceedings of the 39th Annual Cognitive Science Society*, 501–506.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*(3), 307–321.

Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*, e1.

Lieder, F., Griffiths, T. L., Huys, Q. J. M., & Goodman, N. D. (2017). The anchoring bias reflects rational use of cognitive resources. *Psychonomic Bulletin & Review*, *25*(1), 322–349.

Lucas, C. G., Bridgers, S., Griffiths, T. L., & Gopnik, A. (2014). When children are better (or at least more open-minded) learners than adults: Developmental differences in learning the forms of causal relationships. *Cognition*, *131*(2), 284–299.

Lucas, C. G., & Griffiths, T. L. (2010). Learning the form of causal relationships using hierarchical Bayesian models. *Cognitive Science*, *34*(1), 113–147.

Rich, A. S., & Gureckis, T. M. (2018). The limits of learning: Exploration, generalization, and the development of learning traps. *Journal of Experimental Psychology: General*, *147*(11), 1553–1570.

Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, *20*(12), 883–893.

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.

Schulz, L. E., & Sommerville, J. (2006). God does not play dice: Causal determinism and preschoolers' causal inferences. *Child Development*, *77*(2), 427–442.

Smith, C., Carey, S., & Wiser, M. (1985). On differentiation: A case study of the development of the concepts of size, weight, and density. *Cognition*, *21*(3), 177–237.

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.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, *38*(4), 599–637.