Causal intervention strategies change across adolescence

Kate Nussenbaum*, Alexandra O. Cohen*, Zachary J. Davis, David Halpern, Todd M. Gureckis, Catherine A. Hartley
{**katenuss, ali.cohen, zach.davis, david.halpern, todd.gureckis, cate**} **@nyu.edu**
Department of Psychology, New York University
New York, NY 10003 USA

Abstract

Intervening on causal systems can illuminate their underlying structures. Past work has shown that, relative to adults, young children often make intervention decisions that confirm single hypotheses rather than those that discriminate alternative hypotheses. Here, we investigated how the ability to make informative intervention decisions changes across development. Ninety participants between the ages of 7 and 25 completed 40 different puzzles in which they had to intervene on various causal systems to determine their underlying structures. We found that the use of discriminatory strategies increased through adolescence and plateaued into adulthood. Our results identify a clear developmental trend in causal reasoning, and highlight the need to expand research on causal learning mechanisms in adolescence.

Keywords: cognitive development; information-seeking; hypothesis testing; causal learning

Introduction

We frequently take actions to manipulate the causal systems that make up our environments. Critically, these causal interventions often vary in the information they reveal (Bramley, Lagnado, & Speekenbrink, 2014; Tong & Koller, 2001; Coenen, Rehder, & Gureckis, 2015).

Imagine, for example, a child tending to a plant. She might believe that the plant requires sunlight, water, and fertilizer to grow. The child might intervene to confirm this hypothesis by placing her plant on a sunny window sill, watering it daily, and fertilizing it. If the plant blooms, she will take this as evidence confirming her initial hypothesis. However, if she were to consider a competing hypothesis – that the plant needs only water and sun but not fertilizer to flourish – she could instead provide the plant with water and sunlight, and critically, withhold fertilizer. If the plant were to wither, she would gain evidence in favor of her first hypothesis, but if it were to grow, she would gain evidence in favor of the second. In this way, different intervention decisions bring about different sets of evidence that help to discriminate competing ideas.

Consistent with this example, previous research has identified two broad classes of decision strategies for making interventions: Confirmatory interventions seek evidence consistent with a particular hypothesis, while discriminatory interventions seek information that can disambiguate competing alternatives (Coenen et al., 2015). It is unclear, however, how causal intervention strategies change with age. Previous work suggests that children as young as 2 years old can derive sophisticated causal knowledge about the structure of their environment by updating their prior assumptions about cause and effect as they encounter new evidence (Gopnik et al., 2004). This evidence is often self-generated – children perform their own "experiments" during play by intervening on causal systems to resolve their uncertainty about how they work (Gopnik, 2012).

Though children are capable of making informative interventions to drive their own learning (Bonawitz, van Schijndel, Friel, & Schulz, 2012; Schulz & Bonawitz, 2007; Sobel & Sommerville, 2010), their information gathering strategies may be sub-optimal. For example, early work in children's hypothesis testing suggests that the ability to systematically test competing alternatives improved from age 5 to age 11, but that even 11-year-olds often failed to make interventions that would enable them to learn underlying causal rules (Rieber, 1969). In a different experiment, when 9- to 11-year-olds were tasked with determining the cause of a specific chemical reaction, the majority of children failed to design systematic experiments that would enable them to efficiently isolate the causal agent (Kuhn & Phelps, 1982).

Characterizing developmental change in causal reasoning

While this work hints that there may be changes in causal intervention strategy across development, no prior work has systematically characterized these changes from childhood to adulthood, perhaps due to the inherent difficulty in measuring developmental change in this complex ability. Multiple strategies can promote effective inference, so studies that have examined only the accuracy of causal judgments, or that have allowed children to freely manipulate causal systems by performing many different actions, may not effectively capture subtle changes in strategy use across development.

A recent study of adults (Coenen et al., 2015) developed a Bayesian measurement model for determining the extent to which confirmatory vs. discriminatory intervention strategies are invoked during decision-making. In this study, adults' intervention decisions were best characterized by a model that combined the discriminatory Expected Information Gain (EIG) strategy with a Positive Testing Strategy (PTS) that assigned "value" to intervention decisions based on the proportion of causal links they would activate. This intervention strategy is generally less cognitively effortful than more discriminatory strategies and can yield informative outcomes in some contexts (Austerweil & Griffiths, 2011), but can also hinder learning by failing to rule out alternative causal models (Nickerson, 1998). Further, adults increased their use of a discriminatory strategy after attempting to solve problems in which confirmatory interventions were systematically less effective, but decreased their discriminatory strategy use under time pressure (Coenen et al., 2015).

The task and modeling approach used by Coenen et al. (2015) has several key properties that make them particularly well-suited to characterize changes in causal intervention strategy across development. First, the task itself is easy to understand but challenging to perform optimally, such that it can be understood by young children while remaining sensitive to changes in causal learning that may occur throughout late childhood, adolescence, and early adulthood. Second, the modeling approach can effectively capture both optimal, discriminatory intervention decisions, but also the more cognitively simple, confirmatory strategy that may be adopted by resource-constrained learners. Finally, the model enables estimation of continuous strategy mixture weights for each participant, which can characterize the extent to which their choices reflect confirmatory or discriminatory strategies. By leveraging this measure, we can both account for heterogeneity in strategy use across individuals and examine how strategy use may change across development.

Two previous studies have taken a similar approach but have only examined the choices made by young children, between the ages of 5 and 8 (McCormack, Bramley, Frosch, Patrick, & Lagnado, 2016; Meng, Bramley, & Xu, 2018). In both these studies, rather than selecting interventions that maximized their ability to disambiguate multiple competing possibilities, children often made choices that maximized positive evidence in favor of a *single* hypothesis. However, these studies used only a small number of trials, potentially leading to unreliable estimates of strategy use and preventing the examination of learning over time.

Further, selecting interventions that maximize information gain may require multiple cognitive mechanisms that continue to develop throughout late childhood and adolescence. When faced with intervention decisions, individuals must prospectively imagine the outcomes that different actions are likely to bring about (Sloman & Lagnado, 2015). Then, they must evaluate whether these outcomes provide evidence for one causal hypothesis over another to ultimately choose which action to take (Coenen & Gureckis, 2015). Finally, individuals need to recognize that this cognitive process is "worth it" – that considering possible outcomes of different interventions promotes more accurate hypothesis evaluation relative to other less effortful cognitive strategies. Each of these component mechanisms undergoes marked change throughout development. The ability to use mental models of the environment to prospectively compare decisions (Decker, Otto, Daw, & Hartley, 2016), the ability to infer causal relations based on observed outcomes (Gopnik et al., 2017), and metacognitive sensitivity to the efficacy of different cognitive strategies (Weil et al., 2013) all improve not just in early childhood – a focal point of many studies of causal learning – but continuously across late childhood, adolescence, and early adulthood.

Here, we leveraged the approach introduced by Coenen et al. (2015) – and its key measurement characteristics – to determine the developmental trajectories of causal learning strategies across late childhood, adolescence and early adulthood. Though these developmental periods have been largely neglected in the causal intervention literature, research focused on related cognitive mechanisms suggest these periods may be characterized by robust change in learning and decision-making strategies. Beyond characterizing the general trajectory of change in the use of different intervention strategies, we sought to illuminate interactions between different cognitive mechanisms that may support the emergence of discriminatory hypothesis testing.

Methods

Participants

Ninety 7-to-25-year-olds ($M_{age} = 15.87$ years, $SD_{age} = 5.26$ years, range = 7.04 - 25.74 years, 46 females) participated in the study. All participants completed the matrix-reasoning and vocabulary section of the Wechsler Abbreviated Scale of Intelligence, from which age-normed IQ scores were derived. There was not a significant relation between age and IQ in our sample, $F(1,88) < .001, p > .99, \eta_p < .001$.

Task

Participants completed a computerized task in which they were told they were employees at a computer chip factory, whose job was to sort 3- and 4-node computer chips based on the configuration of their hidden wires. On each trial, participants first viewed two causal graphs for 2 seconds, each of which displayed a different possible configuration of the chip's hidden wires (Figure 1). Then, a computer chip appeared, with all of its nodes turned "off." Participants had as much time as they wanted to make one intervention decision that is, to click on one node. The node that was clicked *always* turned on. After a brief delay (200 ms) during which the chip turned grey and beeped, the chip reached its final state, indicating outcome of the intervention. The activation of a parent node turned on its direct descendants with with a probability of .8. There were no background causes - nodes could only turn on if they were directly clicked or activated by a parent node. After viewing the outcome of their intervention, participants had unlimited time to click on whichever of the two causal graphs they believed indicated the true configuration of the chip's hidden wires. Participants then used a continuous slider to rate their confidence that they selected the correct configuration. Participants were told that they would be paid a bonus based on how many chips they sorted correctly.

Prior to beginning the experimental trials, all participants completed an extensive tutorial in which they were trained on the probabilistic nature of the wires, the directionality of the wires, the correspondence between the causal graph diagrams and the actual chip on which they intervened, and the overall trial procedure.

Participants completed 40 experimental trials. Trial order was pseudo-randomized such that in each block of 10 trials, participants always completed five 3-node puzzles and five 4-node puzzles. The specific puzzles were selected such that the

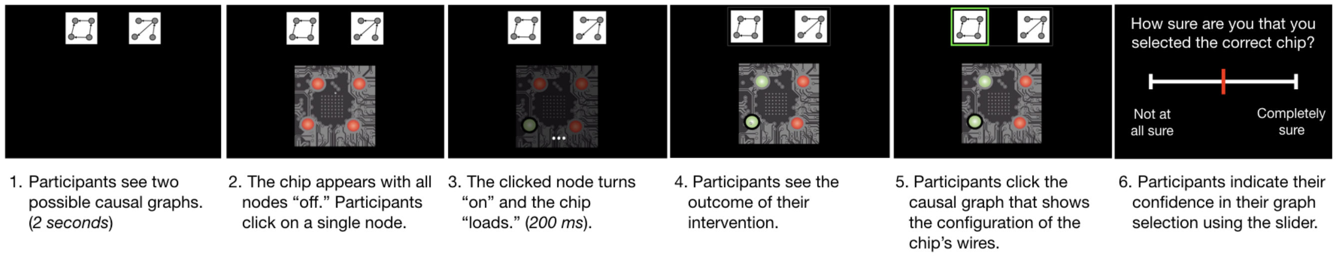

Figure 1: Participants completed 40 intervention trials, in which they had to select a node to determine the configuration of a computer chip's hidden wires.

discriminatory and confirmatory strategy we modeled (more details below) made divergent predictions about the probability of selecting different nodes. The side of the screen on which each graph appeared was randomized. On each trial for each participant, one graph was randomly selected to be the chip's "true" underlying structure. Participants only learned how many chips they sorted correctly at the end of the task; they did not receive trial-by-trial feedback.

Strategies

To model participant intervention choices, we focused on one specific discriminatory intervention strategy - Expected Information Gain - and one specific confirmatory strategy - Positive Testing Strategy. The models differ in how they assign value to possible interventions.

Expected Information Gain (EIG) EIG assumes that individuals have a set of hypotheses about the structure of a particular causal system, with each system represented as a causal Bayesian graph. A learner's uncertainty about which graph (g) is most likely the source of their current observations is represented as the Shannon entropy over the graphs within their hypothesis set (G):

$$H(G) = \sum_{g \in G} P(g) log_2 \frac{1}{P(g)}$$

Learners maximizing information gain should select the information that will cause the largest reduction in their uncertainty. This can be computed by considering the amount of information gained by each possible outcome (o) of each action (a), weighted by their probability:

$$EIG(a) = H(G) - \sum_{o \in O} P(o|a) H(G|a,o)$$

where $H(G|a,o)$ is the new uncertainty after an intervention:

$$H(G|a,o) = \sum_{g \in G} P(g|a,o) log_2 \frac{1}{P(g|a,o)}$$

Positive Testing Strategy (PTS) PTS assumes that participants seek positive evidence to confirm a single hypothesis. We use the formalization introduced in Coenen et al. (2015) which assumes that participants consider each graph in turn, and choose the intervention that will activate the largest proportion of nodes within a single causal graph:

$$PTS(a) = \max_g \left(\frac{DescendantLinks_{n,g}}{TotalLinks_g} \right)$$

Results

Age-related change in strategy use

To characterize participants' intervention choices, we fit a single Bayesian model in which we assumed participants were linearly combining EIG and PTS with weight θ, where $\theta = 0$ indicates a pure PTS strategy and $\theta = 1$ indicates a pure EIG strategy. We further assumed that participants' choices were noisy, such that the expected value of each choice probabilistically influenced intervention decisions. We used a softmax choice function to represent this process, with a free parameter, τ, to capture each participant's decision noise.

The two previous studies using this modeling approach employed a hierarchical model in which group-level hyper-parameters were also estimated (Coenen et al., 2015; Meng et al., 2018), but given our broad age range, we did not want to assume that the participants in our sample comprised a single group. Rather than estimating group-level hyper-parameters, we estimated the model separately for each participant.

We estimated posterior distributions over the parameters using Markov chain Monte Carlo (MCMC) sampling via the NUTS algorithm implemented in STAN (4 chains of 2000 iterations, 1000 per chain discarded as warmup; 4000 total samples per parameter) (Stan Development Team, n.d.; Team, 2013). We used uniform priors over the parameter space ($\tau \sim U(0,\infty); \theta \sim U(0,1)$). Rhat values for all parameter estimates were less than 1.1, indicating convergence across chains (Brooks & Gelman, n.d.).

To characterize how strategy use changed with age, we extracted the posterior mean estimates of strategy mixture weights (θ) and examined their relation with age. We tested two linear regression models to examine linear and nonlinear trajectories of developmental change: One included linear z-scored age as a predictor, and one included both linear z-scored age and quadratic z-scored age as predictors (Somerville et al., 2012). We followed this approach for all subsequent models described in the paper.

The model with the quadratic age term provided a significantly better fit to the data, $F(1,87) = 9.95, p = .002$. Both age ($\beta = .12, p < .001$) and age^2 ($\beta = -.06, p = .002$) significantly predicted strategy mixture weight (Figure 3), suggesting that through early adolescence, participants decreased their use of PTS in favor of EIG. Even within age groups,

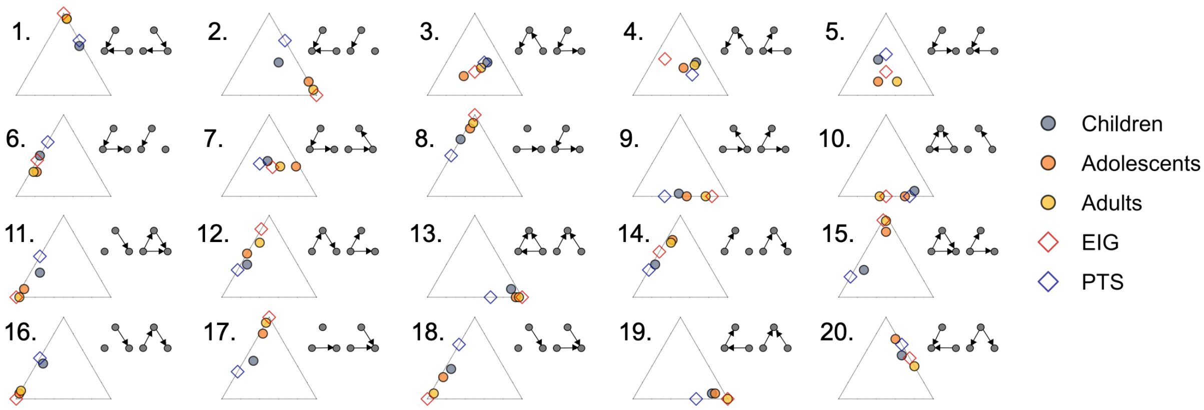

Figure 2: Intervention choices for the 20 three-node puzzles presented in the experiment. The corners of each simplex represent nodes on which participants intervened. The circles represent the average choice for each age group (Children: 7 - 12 years old, Adolescents: 13 - 17, Adults: 18 - 25), while the diamonds represent the "value" of each node as determined by EIG and PTS.

strategy use varied across problems (Figure 2); adolescent choices, for example, sometimes resembled those of adults (16) and sometimes were more like those of children (10).

We also examined how decision noise (τ) changed with age. Decision noise decreased linearly with age ($\beta = -.576, p = .048$), indicating that the choices of older relative to younger participants were more fully captured by the predictions of the two intervention strategies (Figure 3). There was not, however, a significant relation between θ and τ ($p = .271$), suggesting that age-related change in strategy mixture weight can not be attributed to age-related differences in decision noise.

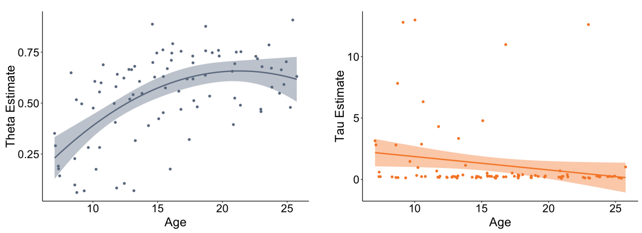

Figure 3: Model-derived estimates of participants' strategy mixture weights (θ) show that participants became more discriminatory with increasing age through late adolescence. Decision noise estimates (τ) show that intervention decisions became more value-based with increasing age. Best-fitting regression lines illustrating the effects of age and age^2 on θ and age on τ are plotted.

In line with previous findings (Meng et al., 2018; Coenen et al., 2015), our modeling results suggest that children and adults use a combination of confirmatory and discriminatory strategies to test causal hypotheses. Further, they demonstrate that this combination systematically differs across children, adolescents, and adults.

Inference-intervention interactions

Why did the use of a discriminatory intervention strategy increase across development? One possibility is that when presented with the novel task, participants explored different intervention strategies until finding one they believed was most effective. Older participants may have been more sensitive to the relative efficacy of different intervention strategies. For EIG to be a useful strategy, however, individuals needed to be able to make accurate causal inferences based on the outcomes of their interventions. Gaining information to disambiguate competing hypotheses was only useful if individuals could correctly updated their beliefs based on that new evidence (Coenen & Gureckis, 2015).

To examine whether causal inference changed with age, we computed the posterior probabilities of each of the two possible causal graphs based on the selected node and the final states of the other nodes on each trial. We then ran a linear mixed-effects model to determine whether there was a relation between age and the posterior probability of the structure selected. Older participants selected more probable causal structures, $F(1,88) = 10.44, p = .002$. This suggests that with increasing age, individuals became better at evaluating the outcomes of their interventions to disambiguate competing hypotheses. However, this metric is inherently confounded with intervention decisions – by definition, interventions with higher EIG scores were more likely to lead to greater increases in the posterior probability of one structure over another. Thus, it is difficult to determine the direction of the relationship between causal intervention and inference – were older participants selecting more informative interventions because they could more effectively prospectively evaluate how that information would enable them to update their beliefs? Or were they updating their beliefs more effectively because they chose interventions that provided stronger evidence in favor of one hypothesis over another?

Participant confidence in the structure they selected can provide insight into developmental change in causal inference – and metacognitive sensitivity to causal evidence – without being confounded by intervention choice. If participants were sensitive to the extent to which the information they gained allowed resolution of competing hypotheses, then their confidence in the structures they selected should track their posterior probabilities. To determine how these posterior probabilities and age influenced confidence ratings, we ran a linear mixed-effects model. Our best-fitting model included both a linear and quadratic effect of age. Participants were more confident in their selection when the posterior probability of the structure they selected was higher, $F(1,3535.17) = 353.69, p < .001$ However, this effect was qualified by an age x posterior probability interaction ($F(1,3529.67) = 21.75, p < .001$) as well as by an age^2 x posterior probability interaction ($F(1,3529.76) = 12.83, p < .001$), such that the influence of posterior probabilities on confidence ratings increased throughout childhood and early adolescence. These results indicate that the ability to evaluate the extent to which new information supported causal hypotheses improved non-linearly across development. Importantly, they suggest developmental improvements in causal inference that are separable from improvements in intervention strategy.

We next examined whether developmental change in causal inference influenced intervention strategy. Specifically, we computed the correlation between the posterior probability of the structure selected and confidence ratings for each participant and ran a linear regression to determine whether these values, which we will refer to as "evidence sensitivity," predicted strategy mixture weight (θ). We found a positive relationship between evidence sensitivity and θ ($\beta = .09, p < .001$), even when controlling for age and age^2. In other words, participants with stronger sensitivity to the strength of the evidence on which to base their inferences also demonstrated greater use of EIG.

Within-task learning effects

Beyond examining how causal intervention strategy changed with age, our use of 40 trials enabled us to examine learning over the course of the task. We hypothesized that older participants' greater use of a discriminatory strategy might in part be driven by faster learning, such that age would more strongly influence estimated values of θ in the second half of the experiment, after participants had the opportunity to learn to adjust their strategy based on their evaluations of their earlier decisions.

To examine whether participants used a different mixture of strategies throughout the course of the task, we fit our Bayesian model separately to the first and second half of trial data for each participant. We then ran a linear mixed-effects model to determine how experiment half and age influenced strategy mixture weight. As before, both linear and quadratic age predicted strategy mixture weight (*ps* $< .02$). Furthermore, strategy mixture weight increased from the first half to the second half of the experiment, $F(1,87) = 11.4, p < .001$ (Figure 4), indicating that participants may have learned to use a more discriminatory strategy over the course of the task. Contrary to our prediction, however, experiment half did not interact with age or age^2 (*ps* $> .20$).

Decision noise also decreased over the course of the experiment, $F(1,88) = 5.18, p = .03$. This effect was qualified by an age x experiment half interaction, such that younger participants demonstrated a greater decrease in decision noise from the first to the second half of trials, $F(1,88) = 4.72, p = .03$ (Figure 4). This suggests that younger participants may have learned to use their estimates of the value of each intervention to more strongly guide their decisions over the course of the task. While the change in their strategy mixture weight did not statistically differ from that of older participants, younger participants may have learned that *both* strategies were more effective than randomly selecting nodes.

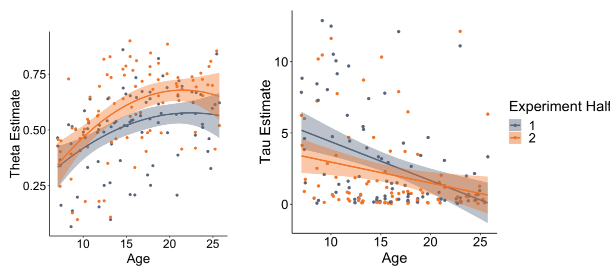

Figure 4: In the second half of the experiment, participants relied more on EIG over PTS, and their choices were less noisy.

Finally, we examined whether evidence sensitivity related to participants' change in strategy use over the course of the task. We computed $\Delta\theta$ for each participant by subtracting their estimated θ value over the first half of the experiment from their estimated θ value over the second half of the experiment. We then ran a regression examining the effects of age and evidence sensitivity on $\Delta\theta$. We found a significant effect of evidence sensitivity on $\Delta\theta$, $\beta = .049, p = .019$, such that participants who were most sensitive to their ability to correctly identify underlying causal structures demonstrated increased use of EIG over the course of the experiment. Mirroring our previously reported results, there was not a significant effect of age on $\Delta\theta$, nor was there an age x evidence sensitivity interaction effect (*ps* $> .61$).

Discussion

Our results and modeling analyses demonstrate robust changes in causal intervention strategy from middle childhood to adulthood. In sum, interventions become more discriminatory with increasing age until reaching a plateau in late adolescence. What causes this developmental shift?

One possibility is that improvements in intervention strategy are due to increased exposure to scientific reasoning strategies through formal schooling. Future work could test participants at multiple time-points and examine the extent to which increases in EIG use align with exposure to curricular units focused on concepts like controlling variables to effectively discriminate hypotheses (Kuhn, Arvidsson, Lesperance, & Corprew, 2017).

However, several aspects of our data suggest that formal schooling can not account for all age-related change in strategy use that we observed. First, almost all participants demonstrated a mixture of strategies throughout the experiment, and this mixture appears to change *gradually* with increasing age (as opposed to a sharp shift corresponding to the introduction of specific concepts during formal schooling). We also found that individual and developmental differences in more basic learning mechanisms, like sensitivity to the informativeness of intervention outcomes, predicted strategy use. Additionally, individuals across our age range increased their use of a discriminatory strategy throughout the course of the task, without any explicit instruction or feedback.

It may also be the case that with increasing age, individuals become better at prospectively planning their intervention decisions. Though evidence sensitivity correlated with strategy mixture weight in our data, it did not fully account for developmental change in strategy use. Importantly, we hypothesized that the ability to make accurate causal judgments may enable individuals to select the best intervention only if they prospectively simulate and sample the outcomes of potential choices in the first place (Bonawitz, Denison, Griffiths, & Gopnik, 2014). On some trials, participants may not have attempted to think through the possible outcomes of their decisions, in which case the ability to evaluate those outcomes would not affect the intervention choice. Future studies should probe the role of other cognitive mechanisms in supporting the use of EIG, like model-based decision-making, which may support or similarly rely on simulating probabilistic outcomes of multi-stage decisions (Decker et al., 2016; Doll, Duncan, Simon, Shohamy, & Daw, 2015).

Another possibility is that younger people are equally *capable* of implementing a more discriminatory intervention strategy, but perform a different cost-benefit analysis when determining which strategy to use. As mentioned previously, the confirmatory PTS strategy often reveals diagnostic information in environments in which causal links are sparse or deterministic (Austerweil & Griffiths, 2011). Additionally, confirmatory hypothesis testing may be adaptive when individuals have the opportunity to make multiple interventions at low cost. It may be the case that rather than spending time and cognitive effort to make the single best intervention, children prefer to make multiple, easier, intervention decisions, which together provide the information they need. Future studies could isolate changes in *ability* from changes in effort allocation, by raising the cost of making an uninformative intervention or forcing all participants to spend a long time deliberating prior to allowing them to perform their intervention.

Finally, though few studies have examined causal learning in adolescence, our results demonstrate that causal learning and decision-making continue to change during this period. Future work probing the cognitive mechanisms that drive these changes will inform how to best support adolescents as they interact with their environments with increasing independence and shape their own learning opportunities.

Acknowledgments

We thank Morgan Glover, Sree Panuganti, Dhiraj Patel, Haniyyah Sardar, Xinxu Shen, and Daphne Valencia for help with data collection. We also thank the Jacobs Foundation (Early Career Fellowship to C.A.H.), the Department of Defense (NDSEG Fellowship to K.N.), the National Science Foundation (Grant No. 1714321 to A.O.C, CAREER grant 1654393 to C.A.H., and CAREER grant BCS-1255538 to T.M.G.), and the James S. McDonnell Foundation (Scholar Award to T.M.G.) for financial support.

References

Austerweil, J. L., & Griffiths, T. L. (2011). Seeking Confirmation Is Rational for Deterministic Hypotheses. *Cognitive Science*, *35*(3), 499–526.

Bonawitz, E., Denison, S., Griffiths, T. L., & Gopnik, A. (2014). Probabilistic models, learning algorithms, and response variability: sampling in cognitive development. *Trends in Cognitive Sciences*, *18*(10), 497–500.

Bonawitz, E., van Schijndel, T. J. P., Friel, D., & Schulz, L. (2012). Children balance theories and evidence in exploration, explanation, and learning. *Cognitive Psychology*, *64*(4), 215–234.

Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (2014). Conservative forgetful scholars: How people learn causal structure through sequences of interventions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 708–731.

Brooks, S. P., & Gelman, A. (n.d.). General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics*, *7*, 434. doi: 10.2307/1390675

Coenen, A., & Gureckis, T. M. (2015). Are biases when making causal interventions related to biases in belief updating? In *Proceedings of the 37th annual conference of the cognitive science society.*

Coenen, A., Rehder, B., & Gureckis, T. M. (2015). Strategies to intervene on causal systems are adaptively selected. *Cognitive Psychology*, *79*, 102–133.

Decker, J. H., Otto, A. R., Daw, N. D., & Hartley, C. A. (2016). From Creatures of Habit to Goal-Directed Learners. *Psychological Science*, *27*(6), 848–858.

Doll, B. B., Duncan, K. D., Simon, D. A., Shohamy, D., & Daw, N. D. (2015). Model-based choices involve prospective neural activity. *Nature Neuroscience*, *18*, 767–772.

Gopnik, A. (2012). Scientific Thinking in Young Children: Theoretical Advances, Empirical Research, and Policy Implications. *Science*, *337*, 1623–1627.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A Theory of Causal Learning in Children: Causal Maps and Bayes Nets. *Psychological Review*, *111*(1), 3–32.

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.

Kuhn, D., Arvidsson, T. S., Lesperance, R., & Corprew, R. (2017). Can Engaging in Science Practices Promote Deep Understanding of Them? *Science Education*, *101*, 232–250. doi: 10.1002/sce.21263

Kuhn, D., & Phelps, E. (1982). Advances in Child Development and Behavior. , *17*, 1–44.

McCormack, T., Bramley, N., Frosch, C., Patrick, F., & Lagnado, D. (2016). Children's use of interventions to learn causal structure. *Journal of Experimental Child Psychology*, *141*, 1–22.

Meng, Y., Bramley, N., & Xu, F. (2018). Children's Causal Interventions Combine Discrimination and Confirmation. In *40th annual meeting of the cognitive science society* (pp. 281–286).

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220.

Rieber, M. (1969). Hypothesis testing in children as a function of age. *Developmental Psychology*, *1*, 389.

Schulz, L. E., & Bonawitz, E. B. (2007). Serious fun: preschoolers engage in more exploratory play when evidence is confounded. *Developmental Psychology*, *43*(4), 1045–1050.

Sloman, S. A., & Lagnado, D. (2015). Causality in thought. *Annual Review of Psychology*, *66*(1), 223-247.

Sobel, D. M., & Sommerville, J. A. (2010). The Importance of Discovery in Children's Causal Learning from Interventions. *Frontiers in Psychology*, *1*, 1–7.

Somerville, L. H., Jones, R. M., Ruberry, E. J., Dyke, J. P., Glover, G., & Casey, B. J. (2012). The Medial Prefrontal Cortex and the Emergence of Self-Conscious Emotion in Adolescence. *Psychological Science*, *24*, 1554–1562.

Stan Development Team. (n.d.). RStan: the R interface to Stan.

Team, R. C. (2013). R: A language and environment for statistical computing.

Tong, S., & Koller, D. (2001). Active learning for structure in bayesian networks. In *Proceedings of the 17th international joint conference on artificial intelligence.*

Weil, L. G., Fleming, S. M., Dumontheil, I., Kilford, E. J., Weil, R. S., Rees, G., ... Blakemore, S.-J. (2013). The development of metacognitive ability in adolescence. *Consciousness and Cognition*, *22*(1), 264–271.