Explaining for the Best Intervention

by

Yuan Meng

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Psychology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Fei Xu, Chair
Professor Alison Gopnik
Professor Mahesh Srinivasan
Professor Neil R. Bramley

Summer 2022

Abstract

Explaining for the Best Intervention

by

Yuan Meng

Doctor of Philosophy in Psychology

University of California, Berkeley

Professor Fei Xu, Chair

Humans don't grow a mind by sitting in an armchair and reading flashcards containing world facts. Much of human knowledge comes from experimentation. For instance, do antidepressants affect both mood and thoughts directly, or do they affect thoughts via mood? To test competing hypotheses, one can *intervene* on a variable (e.g., changing mood using another method) to see how it affects other variables. Good interventions generate *information* that discriminates between hypotheses. However, informative interventions are hard to design. In many past studies, explaining *why* something occurs is used as a simple but powerful tool to help learners acquire generalizable abstractions useful for future scenarios. In this dissertation, I investigate whether asking causal learners to explain why they plan to carry out certain interventions helps them select more informative interventions. Chapter 1 describes the task used throughout the dissertation: Three light bulbs are connected one way or another; learners intervene on one light bulb to find out their true structure. The optimal intervention maximizes the *expected information gain* (EIG) by generating distinct outcomes under different structures. The suboptimal *positive test strategy* (PTS) tests one hypothesis at a time and favors the intervention that can potentially affect the highest proportion of hypothesized connections. A Bayesian model captures how much a learner relies on EIG *vs.* PTS to choose interventions. In Chapter 2 (Study 1), I examine intervention strategies that adults and 5- to 7-year-olds naturally use to select interventions in the Light Bulb Game described above. Adults mainly relied on the optimal strategy, EIG maximization, whereas children mostly used PTS. Following informative interventions, adults identified the correct structures most of the time, but children were at chance. In Chapter 3 (Study 2), I prompt adults and 5- to 7-year-olds to explain their intervention choices (*"Why do you wanna turn on X light bulb?"*) and examine if it changes their intervention strategies. Explainers did *not* intervene differently. However, children who either explained or reported their choices performed above chance at identifying true causal structures from intervention outcomes. In Chapter 4 (Study 3), I train 7- to 11-year-olds on the *difference-making principle* which underlies EIG maximization: That a light bulb is helpful if it makes different things happen in different structures, and unhelpful if it leads to the same outcome either way. Training led children to rely more on EIG. The effect was more pronounced in 9- to 11-year-olds than it was in 7- to 8-year-olds. In Chapter 5, I synthesize findings in Studies 1–3 and propose directions for future research.

To my dad, Weidong Meng.

Contents

Acknowledgments

Growing up as a shy child, I always felt like an observer of human behavior, an outsider, rather than a participant. Finding *Gödel, Escher, Bach* on my dad's bookshelf and reading it at a young age gave me a certain comfort, that for all its complexities and paradoxes, the human mind is perhaps understandable through some kind of mathematical formulations. Years later, when Fei came to Beijing Normal University to give a lecture series, I learned for the first time about computational cognitive science that does just that (if you will allow the gross simplification). I'm grateful that Fei took me on this ride: Using computational modeling to study the human mind at Berkeley is my childhood dream coming true. Years before then, I'm grateful to have met my undergraduate advisor, Qingfen Hu, who opened the door to the wonders of cognitive development for me. Qingfen is also one of the most wholesome and generous human beings I've ever seen, who taught me that being kind is not a transaction you make when you have something to gain, but simply who you are. I was still far too shy when I first came to Berkeley. It wasn't until the end of grad school that I found my voice: Writing this dissertation felt like my coming of age in many ways.

I also thank for my committee members — Alison Gopnik, Mahesh Srinivasan, and Neil Bramley — for the practical advice and insights you gave me every time we met. Berkeley was and is also home to many other academic heroes of mine: Tom Griffiths, Tania Lombrozo, Celeste Kidd, Steve Piantadosi, Anne Collins, to name a few. I wish we had spoken more, but I'm thankful that I got to learn about your fascinating work that sheds light on what human cognition is capable of.

"It takes a village to raise a child" — sometimes, there may not be a village at a particular point in time. I thank past myself for sticking it out. First week into Tom's computational modeling class in Fall 2016, I confided in a fellow student that I had no idea what was going on. That semester, I borrowed Tom's old MATLAB code to complete my final project (later published in the Proceedings of the 39th Annual Conference of the Cognitive Science Society), of which I only had a "computational-level" understanding. I didn't anticipate that for nearly two years since then, I might be the only one left in the Developmental area working on computational models of higher-level cognition. I learned to implement Bayesian cognitive models after various false starts and cycling through a dozen of Bayesian statistics textbooks. If I could go back in time, I would have asked for more help; I won't choose solitary work over fruitful collaborations in my future career again. But I'm grateful for what the struggle has shown me about what I can do in my own time of need.

Many research assistants and summer interns have helped me on this journey. I extend my sincerest gratitude to Anqi Li, Eliza Huang, Sophie Peeler, Shengyi Wu, Jessamine Li, Selena Cheng, Chelsea Leung, Nina Li, Stella Rue, among others. Anqi was my first RA, who taught me a great deal about working with children when I just started. Eliza was my first intern and I'm still amazed by her passion, intelligence, energy, and kindness. After an exhausting day testing at the Lawrence Hall of Science, she would still ask me a ton about science or suggest brilliant ideas for my work.

Much to my surprise, Shengyi was still testing a child on Chinese New Years Eve. This dissertation wouldn't have gotten anywhere without their dedication, which I shall never take for granted. I also thank my labmates for countless thought-provoking discussions and useful suggestions over the years: Rongzhi Liu, Ruthe Foushee, Shaun O'Grady, Elena Luchkina, Stephanie Alderete, Roya Baharloo, Kaitlyn Tang, Emma Roth, Gwyneth Heuser, Harmonie Strohl, and Phyllis Lun.

The best decision I made in grad school was moving to Palo Alto at the beginning of the COVID pandemic. I have since made many new friends and reconnected with old friends in the tech industry and found where I belong, likely for the first time in life. Yishu — thank you for offering advice and suggestions (fresh out of the surgery room, later I learned) when I thought I was going to drop out of grad school, and for telling me how you already thought of me as a data scientist when plenty tried to discourage me, *"Every CS kid wants this job. Why can you get it?"* Wanqiao — you were the first data scientist to lend me a hand on a strange project (one that I still can't make sense of) and let me return the favor later, which made me realize data science is not so much about fancy modeling as it is about common sense, judgment, and camaraderie. Emma — thank you for making me feel valued and seen, and for the many insightful conversations about career, data science, and life. My internship mentor at Bombora, Nishan — thank you for trusting me with such a cool and challenging project, and for your timely encouragement and skillful guidance. Lingyan, Maggie, and Siyu — thank you for advocating for me whenever there was a good job opportunity, offering helpful tips when I was interviewing, and almost lending me money after hearing about my hardships at home. Kefan, Chicheng, Mingxi, and Larry — thank you for the good games, dad jokes, random conversations, and letting me know that you care[*]. Haley — thank you for introducing me to the art of ice-ream eating, and more seriously, for your perspective and open-mindedness. My fascination with cognitive science began with the wish and will to understand the human mind through the lens of mathematics, and it concluded with the wish and will to just be in this moment and enjoy.

In the United States, only one in two PhD students completes grad school. I certainly wouldn't have seen the finishing line had Silvia Bunge not told me, *"You're our student and we'll do all we can to help you."* The only time I cried in the past 6 years, it was for her kindness. I thank Silvia and Julie Aranda for helping me out at the end of 2020 and making it possible for me to finish.

Last but not the least, I thank the financial support from the Berkeley Fellowship for Graduate Study in 2016–2018 and my family over the years. As a child, my parents and I didn't really speak much — frankly, that was why I became intrigued by developmental psychology as an undergrad: I was curious how that relationship is like for somebody else. In the past year and a half, my dad was diagnosed with ALS, after which his health rapidly deteriorated. It wasn't until then that I realized how much impact he had on me: It was him who taught me how to write when I was in middle school, and writing is still one of the things I love the most; the many books he brought home for me ignited my interests in science, history, and philosophy, which last to this day. When I left home 6 years ago, I never thought there would come this day when I dedicate this dissertation to him.

[*] When I finally get to settle down, maybe we can write that Hollywood script about last year.

Listing of figures

Listing of tables

You insist that there is something a machine cannot do. If you will tell me precisely what it is that a machine cannot do, then I can always make a machine which will do just that!

John von Neumann (1948)

Chapter 1

Introduction

Learning is a miracle, with theoretical impossibility contrasted by our apparent competence and efficiency. As a child, I used to worry how one lifetime was ever enough for us to understand a single grain of sand and its infinite facets, let alone the vastness of the universe. Years later in college, I learned that W. V. O. Quine even went so far to argue all human knowledge is epistemically equivalent to Greek mythology, for any theory is underdetermined by empirical data[45,93]. Yet in reality, learning is far from a hopeless endeavor: Before celebrating their first birthdays, human infants have already uttered the first words in their native languages[128] and demonstrated intuitive understandings of physics[4], causation[31,72], agents[5,76], number[33,41,139], probability[35,36], etc..

How do we learn so much from so little so quickly? This is the holy grail of intelligence, and a hard question to answer, for even the very concept of learning is ill-defined (see Schulte[110], for a review of formal learning theories). For instance, can we say someone has learned Go after being taught the rules of Go? Do we learn anything new from deduction given that the conclusion is determined by the premises? Do we know a technology (e.g., bikes or bitcoin) if we can use it but don't understand its inner workings[106]? If understanding is not required for knowing, how often do we differ from the machine in Searle's Chinese room that translates Chinese to English based on an instruction manual[61,114]? To a great extent, modern cognitive science often bypasses epistemic debates about learning and takes up a practical interest: How do humans grow from the seeming "blooming, buzzing confusion"[62] in infancy to competently seeing, thinking, and acting in adulthood?

The most studied form of learning is learning from observation: Human infants are endowed with certain amount of core knowledge[19] as well as powerful inductive biases and inferential mechanisms they can use to quickly extract generalizable knowledge from limited observations[46,52,66,125,138]. As the one of the "Godfathers of AI" Geoffrey Hinton put it, we can learn much by looking alone:

> "When we're learning to see, nobody's telling us what the right answers are — we just look. [...] And there's only one place you can get that much information — from the

> input itself." (cited in Murphy[87], p. 10)

However, human learners do more than just observing. Real-world observations are almost always insufficient or ambiguous — people can run *experiments* collect more data. Even in the absence of new empirical data, we can gain new insights by *thinking* — such as the jury deliberating on evidence presented during the trial to reach a verdict that they believe is closer to the truth than not. Thinking can also take the form of thought experiments, mental simulation, and self-explaining. Experimentation and thinking allow us to generate more data or do more with less. In the next three sections, I first review the state of the art in the three forms of learning as mentioned: Learning by observation, learning by experimentation, and learning by thinking. After reviewing the literature, I propose the leading question in my dissertation: *If thinking facilitates learning from existing data, can it also help learners design more informative experiments to collect useful data in the first place?* To end Chapter 1, I delineate the experimental and formal approaches to addressing this question.

1.1 Learning by Observation

Among various forms of learning, learning by observation has received the most attention. Word learning is a prominent example often used to showcase our remarkable ability to learn from sparse observations[8,18,79]. In Quine's famous "gavagai" problem[101,102], a rabbit runs by as a speaker of an unknown language utters, *"Gavagai!"* You and I might immediately guess that "gavagai" refers to a rabbit, yet this strong intuition is unjustified. Why can't "gavagai" refer to any animal, running, a rabbit minus a leg, a rabbit plus the grass beneath, or countless other possibilities? In the movie *Arrival*, linguist Dr. Louise Banks was faced with this exact challenge when she was hired to translate an alien language she initially knew nothing of. Her once-in-a-generation feat is what human children achieve on a daily basis. For instance, upon observing a novel word "blicket" being paired with several examples of dalmatians, 3- to 4-year-olds selectively generalize "blicket" to other dalmatians, but not other breeds of dogs or animals in general[140,141]. Such efficiency and constraint may be in part due to children's inductive biases for word meaning (e.g., that nouns typically refer to whole objects, not parts[80]) and in part a natural consequence of Bayesian inference (e.g., if "blicket" refers to animals, it would be a "suspicious coincidence" that the speaker uses three dalmatians as examples).

Causal learning is another powerful demonstration of obtaining abstract knowledge from observation[66,97]. Hume was skeptical as to whether causation can be inferred from observation at all[60],

> "When we look about us towards external objects, and consider the operation of causes, we are never able, in a single instance, to discover any power or necessary connexion; any quality, which binds the effect to the cause, and renders the one an infallible consequence of the other. We only find, that the one does actually, in fact, follow the other." (*An Enquiry Concerning Human Understanding*, p. 46)

With metaphysical skepticism set aside, claiming a variable *causes* another is a daunting statistical challenge, since a myriad factors could be at play. Guido Imbens, Joshua Angrist, and David Card won the 2021 Nobel Prize in Economic Sciences for inventing techniques that lend more confidence to such inferences[1]. Yet, human toddlers can infer an object's causal efficacy after observing it covary with the effect a few times[50,51,119]. In Sobel et al.'s study[119], 2- to 3-year-olds saw objects A and B activate a machine together and then A activate the machine alone. Most inferred A as causally efficacious but not B, even though there was no negative evidence against B. This "backward blocking" behavior demonstrated by toddlers aligns with predictions of formal models of causal inference[95,120]. As with word learning, causal learning is also aided by our prior knowledge (e.g., it's more likely that a single cause leads to the effect, rather than multiple causes combined[130]; an effect usually occur shortly after the cause[13]) along with powerful mechanisms to extract abstract knowledge from concrete data (e.g., SARS-CoV-2 causes COVID-19 → viruses cause diseases[56,48]).

In a variety of other domains such relational reasoning (e.g., whether multiple objects are the same or different on certain dimension[37,133]), social learning (e.g., the "essence" or defining features of social groups[105], group affiliations and social structure[47,127]), and intuitive theories of science (e.g., when magnets repel and attract one another[10]), young children have also demonstrated extraordinary abilities to learn and generalize from just a handful of observations. Owing to rapid advances in AI, now is perhaps the golden age of learning by observation. Even without inductive biases or prior knowledge that constrain and speed up human learning, large neural networks with hundreds of billions of parameters trained on huge amounts of observations, or so-called "foundation models"[9], begin to show human-level competence in domains such as language[17,24], vision[38,103], social cognition[7], *etc.*. An astounding example representing the state of the art is DeepMind's Gato, a "generalist agent" that not only can play games, caption images, hold conversations, but can even control a robot arm, all relying on a single transformer model trained on text and image sequences.

1.2 Learning by Experimentation

Where observation ends is where inquiry starts. From Piaget[99] to contemporary cognitive scientists[49,109], children have long been likened to scientists: Both organize their knowledge of the world in the form of formal or intuitive theories, and both gather data to test theories and revise theories in light of data. Children don't sit and watch parents play with toys to learn how toys works, much like scientists who don't just listen to talks to advance science — both *inquire* into mysteries, ask questions, explore around, and run controlled experiments to gain trustworthy causal knowledge.

Granted, children are not literal scientists. In a loose sense of the child-as-scientist metaphor, children know when to seek more data and can learn from self-generated data[25,49,111]. For instance, children don't just explore to gain novel experience, but more so to gain new information. In Schulz et al.'s seminal work[112], one group of preschoolers were given confounded evidence that two levers together activated a machine, whereas other groups observed each lever individually. Choosing between the old or a new toy, children who received confounded evidence explored the old one more, but not those who received clear evidence. Apart from statistical cues, children also use social cues to guide exploration. One example is that if a teacher was found to omit some functions when demonstrating what a toy could do, children explored a new toy shown by this teacher more than if the teacher didn't omit information[57]. From data gathered via exploration, children can learn abstract rules (e.g., two blocks of the same shape/color make the machine go[117]) and find what they are searching for (e.g., where a target is hidden among several locations[107]), to name a few domains.

To be like scientists in a stronger sense, children not only need to seek evidence and learn form it, but the evidence they seek should be maximally useful — that is, child scientists should follow the *optimal experiment design* (OED[28,40,74,89,118]) intuitively. This is a much taller bar to meet. To begin, let's formalize what scientists should do so we can measure children's performance against it.

A classic view in science is that we should isolate variables and manipulate one at a time, so we can attribute the potential effect to the manipulated variable alone[20,65,100]. For instance, if we don't know what determines how far a ball rolls down from a ramp, we can build two ramps that only differ in one aspect, such as height, slope, friction, *etc.*, but not multiple aspects at once. While this Control-of-Variables (CV) strategy works slowly but surely, more efficient strategies are available in many cases, such as simultaneously testing multiple variables when causes are sparse (known as the rarity assumption[3,88,92]). For example, if only one in 20 switches can turn on the light, why not turn on half of the switches at once, do the same to whichever half that turned on the light, and repeat the bisecting process until you find the switch that works? At worst, you finish in five rounds (ruling out 10, 5, 2, 1, and 1 each time). If you try one switch at a time, then it may take as many as 19 rounds (if none of the first 19 turned on the light). In several recent studies, increasingly more 7- to 13-year-old children[14] and adults[30] chose to test multiple variables at once as the causes became rarer.

Economic factors aside, modern OED largely hinges on maximizing the *information* you get from an experiment. To grasp this idea, consider a common problem guitarists face: Pedals creating sound effects can be linked together on a pedal board in any arbitrary order. While on tour, you forgot how the four pedals on your pedal board are connected and the cables are hidden by the sound engineer. There are $3^{\binom{4}{2}} = 729$ possible ways. Fortunately, your notes say they can only be connected in one of three ways, as shown in Figure 1.1. These are called acyclic directed graphs (DAGs) or Bayesian networks[95,120], where nodes represent causal variables and edges show their connections.

To figure out which graph is true, you can turn on a pedal and listen to the effects — fixing the value of a variable to see how it impacts other variables is called *intervention*[95], which is an effective way to learn causal structures. All interventions are not equally useful; to evaluate their usefulness, the concept of information comes into play. Among all graphs $\mathcal{G} = \{g_1, g_2, \ldots, g_n\}$, some (e.g., connected in a chain) are more likely than others (e.g., one pedal linked to all the other). The inverse probability of a graph $1/P(g_i)$ quantifies our surprise of seeing it. Assuming graphs are all independent, their total surprise is $\prod_{i=1}^{n} 1/P(g_i)$. To avoid computing the product, we can use logarithms. $\mathrm{I}(g_i) = -\log P(g_i)$ is called *surprisal.* The expected surprisal of the set is its *information entropy*[115],

$$\mathrm{H}(G) = \sum_{i=1}^{n} P(g_i)\mathrm{I}(g_i) = -\sum_{i=1}^{n} P(g_i)\log P(g_i), \tag{1.1}$$

which encodes our uncertainty about which graph is true before any intervention.

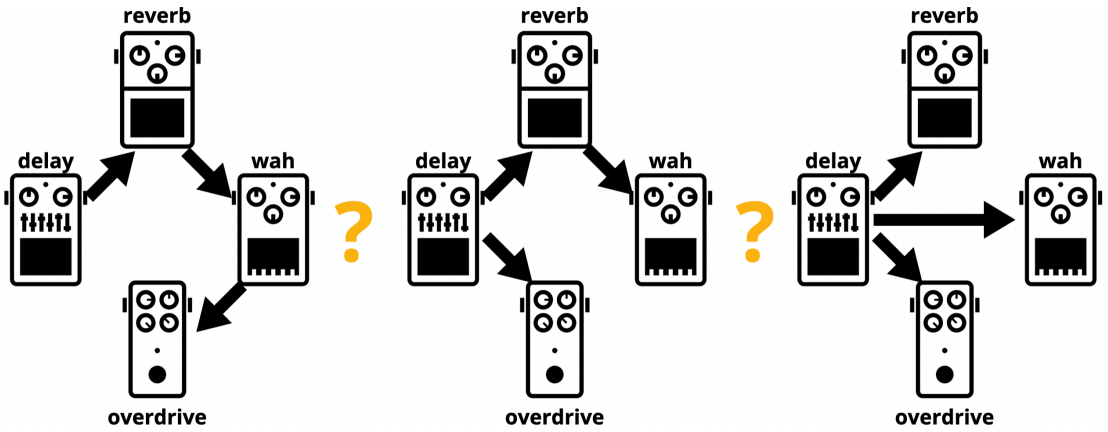


Figure 1.1: A pedal board with delay, reverb, overdrive, and wah pedals connected in an unknown way.

Now, you connect your guitar to the wah pedal and hear no other effects than just "wah". Given this intervention a and its outcome y, the conditional entropy becomes

$$\mathrm{H}(G|y,a) = -\sum_{i=1}^{n} P(g_i|y,a) \log P(g_i|y,a). \tag{1.2}$$

$\mathrm{H}(G|y,a)$ is smaller than $\mathrm{H}(G)$ since you can rule out the first graph; otherwise you'll also hear the overdrive. The reduction in entropy, $\mathrm{H}(G) - \mathrm{H}(G|y,a)$, is called mutual information, or more intuitively, *information gain* (IG). Before carrying out any intervention, however, we cannot foresee what the outcome would be, so IG cannot guide our choice. The best we can do is to maximize the *expected information gain* (EIG) over all potential outcomes $\mathcal{Y} = \{y_1, y_2, \ldots, y_m\}$,

$$\mathrm{EIG}(a) = \mathrm{H}(G) - \sum_{j=1}^{m} P(y_j|a)\mathrm{H}(G|y_j,a). \tag{1.3}$$

In this case, the reverb pedal is the intervention that maximizes EIG: If you also hear "wah" and overdrive, it's the first structure; if only "wah" and reverb, the second; if just the reverb, it's the last.

When causal structures are simple (e.g., two nodes), even preschoolers can perform informative interventions to distinguish between possible structures[32,69]. In Lapidow et al.'s recent study, 4- to 6-year-olds were told, for instance, that gears A and B were both broken or A was broken but B was not. Children could spin one gear to figure out which was the case. Spinning A does nothing either way, whereas spinning B creates an effect (i.e., B spinning) only in the latter but not in the former case. Most children successfully chose B and identified the correct structure afterward. The authors claimed that their study was the first to demonstrate young children could select and learn from informative interventions. However, learners in this gear task were testing the causal efficacy of *nodes* (e.g., does each node work?), not the structure formed by *edges* (e.g., how are nodes connected?). A large part of science is concerned with the latter: Scientists often wish to find out how variables are related, not just examining each individual variable. Moreover, to find the optimal solution (e.g., intervening on the reverb pedal) to the pedal board problem I just described, learners must consider what each node does in each graph as a whole. Testing one edge at a time also solves this problem eventually, but may take 12 interventions (2 interventions per edge × 6 possible edges) in the worst-case scenario. Since the gear task only had two variables and one edge, it was not possible for us to know whether learners were using the less efficient pairwise test strategy or truly maximizing EIG.

When causal structures get more complex (e.g., three or more nodes) and edges are the hypotheses under examination, even school-aged children and adults don't always maximize EIG. Some learners forget[15] or fail to integrate[122] evidence gathered from past interventions. Some use the pairwise test strategy mentioned above[42] or update their current causal hypothesis one edge at a time in light of data, rather than considering all possible structures at once[12].

Moreover, children and adults[83,85,86,90] often exhit a root preference (or "source preference"): In the pedal board example, this preference would manifest as turning on the delay pedal, which creates all sound effects in all graphs but leaves learners uncertain as to which one is true. McCormack et al.[83] used a similar task to test 5- to 8-year-olds, showing them three ways in which a three-node causal system might work: That one node served as a common cause of the other two ($B \leftarrow A \rightarrow B$) or three nodes worked in a causal chain ($A \rightarrow B \rightarrow C$ or $A \rightarrow C \rightarrow B$). Children were allowed to intervene at least 12 times. The chance performance was simulated by randomly selecting interventions. Only 7- to 8-year-olds' intervention quality was above chance for both the common cause and the causal chain structures. By contrast, 5- to 6-year-olds' intervention quality was below chance for both types; 6- to 7-year-olds' was above chance for causal chains but not for the common cause. The researchers noted that most common mistake was to intervene on the root node A.

Why do learners sometimes intervene on root nodes that may not be informative? Inspired by the rule learning literature[3,63,88,92,134], Coenen et al. hypothesized that these "suboptimal" learners may be using the positive test strategy (PTS) and their goal is to test one hypothesis at a time efficiently. In the pedal example, if you are testing the first graph, you can use the delay pedal to simultaneously test all three links: If *not* all effects are produced, you can rule out this graph. The reverb and the wah pedals are less useful, since they can only test a subset of links in the first graph. The overdrive pedal is unhelpful, as it cannot test any links at all. An intervention's PTS score is thus defined as the proportion of links it can test in a given graph. If an intervention a has different PTS scores in different graphs G, the maximum score across all graphs is assigned as its final score,

$$\mathrm{PTS}(a) = \max_{g}[\frac{DescendantLinks_{a,g}}{TotalLinks_{g}}]. \tag{1.4}$$

The delay pedal that always tests all links has the highest PTS score of 1. However, as mentioned, since the effect is the same across graphs, you need additional tests to identify the correct one. As a result, PTS is less optimal than EIG, which requires exactly one intervention here. From early childhood (5-year-olds[85,86]) to adolescence[90] and adulthood[29], causal learners partly rely on EIG maximization and and partly PTS, with the reliance on EIG gradually increasing with age[90].

1.3 Learning by Thinking

Another major form of learning distinct from learning by observation is learning by thinking (LbT). The hallmark of LbT is that learners can again new insights without new empirical observations[78], by means of thought experiments[26], mental simulation[59,126], or explaining to oneself[22,23,43,77]. Among them, thought experiments have received the most attention from philosophy and science, but are the least studied in cognitive science[78]. This asymmetry may result from the abundance of thought experiments in scientific discoveries (e.g., Schrödinger's cat, Einstein's train, Hawking's turtles, Galileo's balls, Maxwell's demon, to name a few) and the scarcity of them in everyday thinking*.

By contrast, explaining to oneself is far more accessible and commonplace — we all ask ourselves "why?" all the time and constantly answer our own questions — yet this seemingly simple act has profound influences on both scientific and common sense reasoning. A fascinating example in science is Olbers' paradox[58]. Looking at the night sky, we see darkness in between stars. If light has traveled infinitely long, it would have reached everywhere in the universe. So why isn't the night sky full of starlight already? There's an elegant explanation: Light has only traveled for a finite amount of a time, which suggests the universe may have a beginning after all. Explaining two ordinary observations has led to an extraordinary discovery in cosmology. Like scientists, laypeople also benefit from explaining how and why things happen[77,135]. For instance, explaining why a toy plays music led preschoolers to generalize based on internal features (e.g., what's inside the toy, such as batteries) rather than perceptual similarity (e.g., the color of the toy)[132]. Explaining why an exemplar belongs to a category helped adults discover subtle rules that explained all cases, rather than salient rules that only explained most[136]. Explaining why something happened in a story enabled children to grasp the moral the story — when asked to choose a similar story, they identified one that shared the same moral, even though concrete plots and characters differed[131]. Explaining a difficult passage on human biology helped students understood it better (measured in a test) than reading it twice[23].

Why does explaining why facilitate learning? Multiple mechanisms may be at play. Michelene Chi and colleagues argued that explaining helps learners identify gaps in their knowledge, catch incorrect assumptions, repair inconsistent mental models, and integrate new information with existing knowledge[21,22,23,43]. Theoretically, other processes that serve these purposes can similarly facilitate learning. Tania Lombrozo and colleagues proposed a mechanism unique to explaining: All else being equal, we view *simple* and *broad* explanations as better than explanations which invoke unnecessary assumptions or fail to capture some observations; when we engage in explaining, the process recruits the aforementioned "explanatory virtues" (i.e., simplicity and breadth, among others), which constrain our downstream learning and inference. Empirical results corroborated these predictions. Both children[11] and adults[94] favor simple explanations to begin with: If two effects (weight and sleep losses) can be explained by one common cause (e.g., depression), they prefer it to

*When was the last time you thought of a thought experiment? Take myself for example, I can only think of one that I came up with on my own. In a conversation with a friend, I asked if they'd rather live a happy life but die remembering a life full of misery, or live a miserable life but die with happy memories of a life fulfilled. The point of my thought experiment was to get at our intuitions about where the meaning of life stems from, whether it be the factual experience or what we believe the experience to be.

two separate causes (e.g., eating and sleeping disorders). Asking children to explain why the effects occurred further exaggerates their simplicity preference[129]. Interestingly, explaining doesn't always lead to the right inference. As mentioned, explaining can help learners find broad patterns that cover all observations[136], but when no such pattern exists, explainers sometimes ignore anomalies and oversimplify the rule, a finding dubbed as the "hazards of explanation"[137]. Explaining also has other effects on learning, such as encouraging comparison[39] and invoking prior knowledge[133,137].

Learners don't need to find the "Inference to the Best Explanation" (IBE) to reap the benefits of self-explaining — the act of engaging in explaining suffices, a phenomenon Wilkenfeld and Lombrozo call "Explaining for the Best Inference" (EBI). The bottom line is, regardless of the mechanism, explaining often helps learners discover generalizable abstractions useful for future scenarios.

1.4 Explaining for the Best Intervention

At the core of this dissertation is the mystery I laid out in the beginning: *How do people learn so much from so little, when the most powerful machines struggle to have human-like common sense?* To summarize the answers from the literature: Apart from inductive biases and efficient learning algorithms[44,66], human learners can go beyond data fed by others — even children know when to gather more data and can learn from self-generated data and explaining helps people draw more robust inferences from existing data. However, compared to learning by observation, the limits and mechanisms of learning by experimentation and learning by thinking are less well understood.

Designing good experiments is challenging, even for scientists. Optimal experiment design (OED) requires learners to intervene on variables that maximize the expected information gain (EIG) over potential outcomes. Children can only do so when testing simple causal hypotheses that only involve a pair of variables (i.e., whether two nodes can both activate or only one of them[69]). When it's necessarily to consider the global structure of three or more variables, Nussembaum et al. found that children didn't primarily rely on EIG until after middle school (12 years or older). All empirical evidence to my knowledge[83,90], including what I'm going to report in this dissertation[85,86], suggests that children across a wide age range (5 to 11 years old) primarily rely on the suboptimal positive test strategy (PTS) to select interventions, which adults occasionally do as well[27,29].

To advance the field of learning by experimentation, a key question is, *what can we do to help causal learners select better interventions?* A potential answer is asking them to explain why they intervene on certain variables. Explaining promotes generalizable abstractions[77,133] and EIG maximization is supported by a simple abstraction — the "difference-making" principle: Regardless of the specific causal systems you are testing, you should always aim to intervene on variables that lead to distinct outcomes under different structures. Explaining promotes comparisons[39] and EIG maximization requires comparisons at various stages — for each intervention choice, one needs to compare how the outcome differs in each hypothesized structures; when choosing the final intervention, one compares the value of all choices. Apart from a recent study[16], all work on self-explaining that I know of has only focused on how explaining facilitates learning from existing data.

To facilitate learning from experimentation and explore the boundary of learning by thinking, I propose the leading question in this dissertation: *If thinking, in the form of explaining, facilitates learning from existing data, can it help learners design better experiments to collect useful data?*

Several studies have hinted at the connection between explaining and intervention, or more generally, data generation. For instance, when a toy's behavior changed during a study, children's spontaneous explanations of why it happened correlated with how they explored the toy[71]. If children suspected a categorical change, they sorted the toys; if they thought the toy was handled differently, they tried different actions on it. In a recent reinforcement learning (RL) study[67], RL agents trained to generate verbal explanations for why a choice was correct or incorrect learned to intervene on object features to change the outcomes, whereas RL agents not trained to explain didn't learn to intervene. Outside of causal learning, explaining was found to improve children's question asking[108] and encourage adults to observe more before making risky decisions[75]. The evidence was indirect because Legare *didn't* manipulate whether or how children explained to see how it impacted their subsequent exploration. It could be that different children had different beliefs about the toy's behavior, which manifested in both their explanations and exploration. Lampinen et al. did manipulate whether or not learners explained but looked at machine rather than human learners. Studies by Ruggeri et al. as well as Liquin and Lombrozo were not in the causal learning domain.

In this dissertation, I investigate whether and how explaining impacts causal intervention. To answer this question, I take an experimental and computational approach, which I will delineate in the next two sections. The rest of the dissertation is organized as follows: In Chapter 2 (Study 1), I characterize intervention strategies 5- to 7-year-old children and adults naturally use to select interventions. In a causal learning task similar to the pedal example, learners intervene on one of three nodes and use the intervention outcome to identify the correct causal structure among two. I use a Bayesian model adapted from Coenen et al.[29] to capture how much each learner relies on the optimal EIG maximization strategy *vs.* the suboptimal positive test strategy (PTS). In Chapter 3 (Study 2), I prompt 5- to 7-year-olds and adults to explain their intervention choices and examine whether they rely more on EIG. In Chapter 4 (Study 3), I train 7- to 11-year-olds to generate explanations that hinge on the difference-making principle behind EIG maximization to see if they intervene better. In Chapter 5, I synthesize findings in Studies 1–3 and propose directions for future research.

1.5 Computational Modeling

As with past studies[27,29,90], I assume causal learners sometimes use EIG and sometimes PTS. By "using" a strategy, I don't mean that learners literally carry out the required computations. Rather, I focus on what Marr[81] calls the *computational* level and aim to capture the problem learners solve when they select interventions: Is the problem EIG maximization, PTS, or a bit of both? When solving the problem, learners can use approximate algorithms to avoid expensive computations[12,55].

Suppose that a hybrid learner *i*'s weight of EIG is θ_i and weight of PTS $(1 - \theta_i)$, then an intervention's value is the weighted sum of its EIG and PTS scores:

$$V(a) = \theta_i \cdot \text{EIG}(a) + (1 - \theta_i) \cdot \text{PTS}(a). \tag{1.5}$$

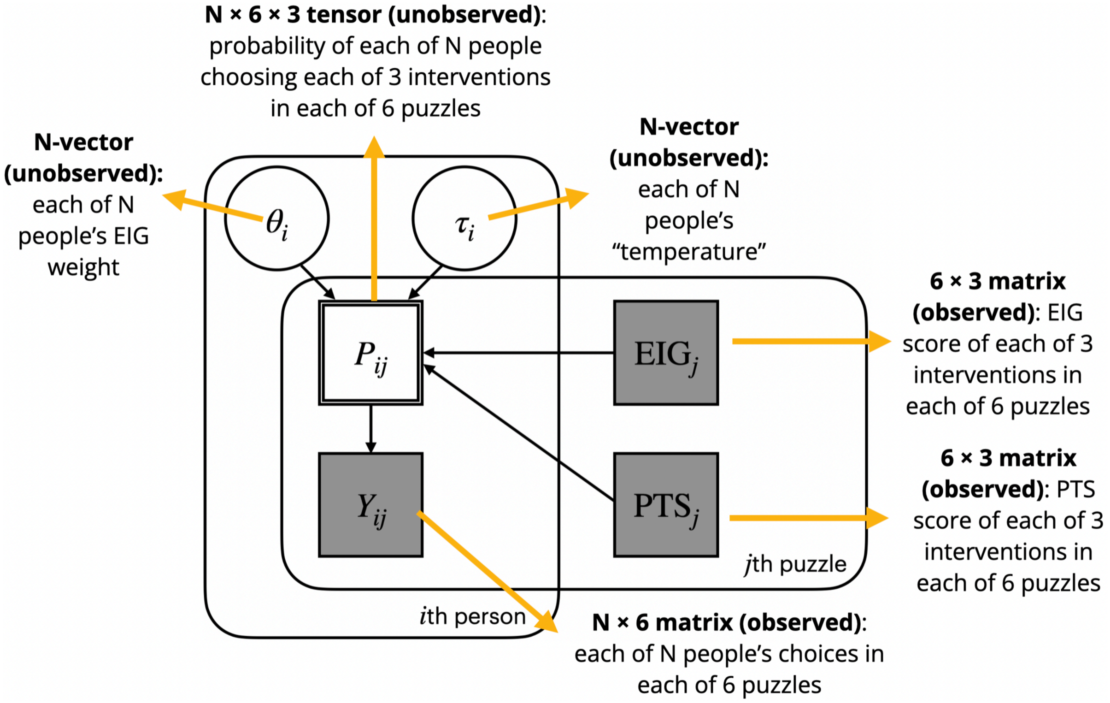

N × 6 × 3 tensor (unobserved): probability of each of N people choosing each of 3 interventions in each of 6 puzzles
N-vector (unobserved): each of N people's EIG weight
N-vector (unobserved): each of N people's "temperature"
θ_i
τ_i
P_{ij}
Y_{ij}
EIG_j
PTS_j
6 × 3 matrix (observed): EIG score of each of 3 interventions in each of 6 puzzles
6 × 3 matrix (observed): PTS score of each of 3 interventions in each of 6 puzzles
jth puzzle
ith person
N × 6 matrix (observed): each of N people's choices in each of 6 puzzles

Figure 1.2: Graphical representation of the Bayesian data analysis model: In each puzzle j, the probabilities of participant i choosing each of the three possible interventions are a 3-vector p_{ij}, which are influenced by the three interventions' EIG_j and PTS_j scores as well as participant i's EIG weight θ_i and decision noise τ_i. Participant i's intervention in puzzle j is chosen from a categorical distribution, $Y_{ij} \sim \text{Categorical}(\mathrm{p}_{ij})$.

More valuable interventions are more likely to be selected. The value of the intervention a, $V(a)$, is linked to the probability of the learner choosing a via a softmax function[123]:

$$\mathrm{P}(a) = \frac{\exp\left(V(a)/\tau_i\right)}{\sum_{a \in A} \exp\left(V(a)/\tau_i\right)}, \tag{1.6}$$

where the temperature parameter captures this learner's decision noise: When $\tau_i = 0$, the learner always choose the intervention with the highest value; when $\tau_i = +\infty$, they intervene randomly.

Figure 1.2 depicts the Bayesian model used to infer each learner's θ_i and τ_i from their intervention choices. Coenen et al.[29] used a hierarchical Bayesian model (HBM), which sampled τ_i and θ_i from group-level hyperparameters. Since participants in my work (adults and children of different ages) are much more heterogeneous than theirs (adults on Amazon Mechanical Turk), I will fit parameters individually for each learner rather than assuming one group sharing the same hyperparameters.

In computational cognitive science, Bayesian models are used in two distinctive ways[124] — as cognitive models to capture how people learn from data and as data analysis models to learn about people from data. The Bayesian model described above is more of a tool for analyzing behavior data.

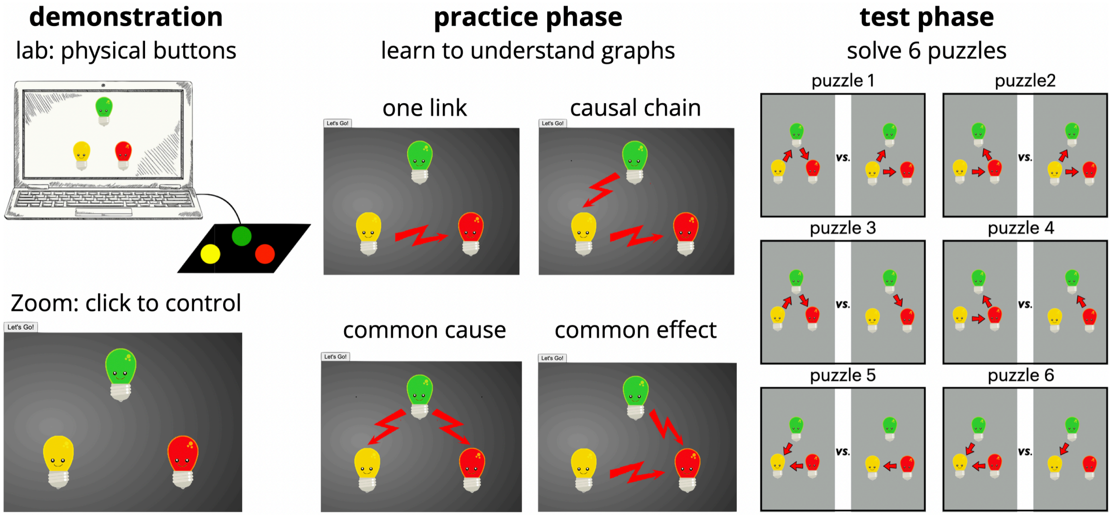


Figure 1.3: The Light Bulb Game consists of three phases: A demonstration phase where the experimenter teaches participants how to control three light bulbs, a practice phase where participants learn to read 4 causal structures, and a test phase where each participant solves 6 puzzles in a randomized order.

1.6 The Light Bulb Game

Each learner's intervention strategy is characterized by their EIG weight θ and PTS weight $(1 - \theta)$. To measure intervention strategies across a wide age range, I adapted the task from Coenen et al.[29] (see also Nussenbaum[90]) to create the Light Bulb Game, where learners can use one intervention to identify an unknown three-variable causal system's true structure (the right panel in Figure 1.3).

Each causal system consists of three light bulbs connected in one of two possible ways. If a light bulb has an arrow pointing to another light bulb, it can turn on the latter. Unlike previous studies where causal relationships were probabilistic[27,29,90,122], I use deterministic relationships to make the game more straightforward: If any cause is present, an effect will 100% occur; if no cause is present, there won't be any effect. To solve a puzzle, learners can only intervene on one light bulb to identify the correct structure. In each puzzle, only one light bulb creates distinct effects in different structures — it has an EIG score of 1; the other two lead to the same outcome in both and therefore have an EIG score of 0. The PTS score of each light bulb can be 0, .5, or 1, depending on the maximum proportion of links it can test between the two structures. Each learner solves six puzzles in a randomized order. Across 18 possible interventions in six puzzles, the mean EIG score is .33 and the mean PTS score is .57. If learners intervene randomly, these are the mean scores we expect to see.

The Light Bulb Game is used in all studies. In Studies 1–2, participants were tested in person on a physical device. In Study 3, participants were tested over Zoom, playing the game on a web browser.

Twenty skillful hypotheses will ascertain what two hundred thousand stupid ones might fail to do.

Charles S. Pierces (1901)

Chapter 2

Natural Intervention Strategies

To begin, what strategies do people naturally use to select interventions? To capture learners' baseline strategies in Study 1, I will test children and adults in the Light Bulb Game (described in Chapter 1) without offering them any help or feedback. For child learners, I chose the age range 5 to 7 years (60 to 95.9 months) based on several past studies in which children selected their own interventions (e.g., 4- to 6-year-olds in Lapidow and Walker[69], 5- to 8-year-olds in McCormack et al.[83,84]).

2.1 Method

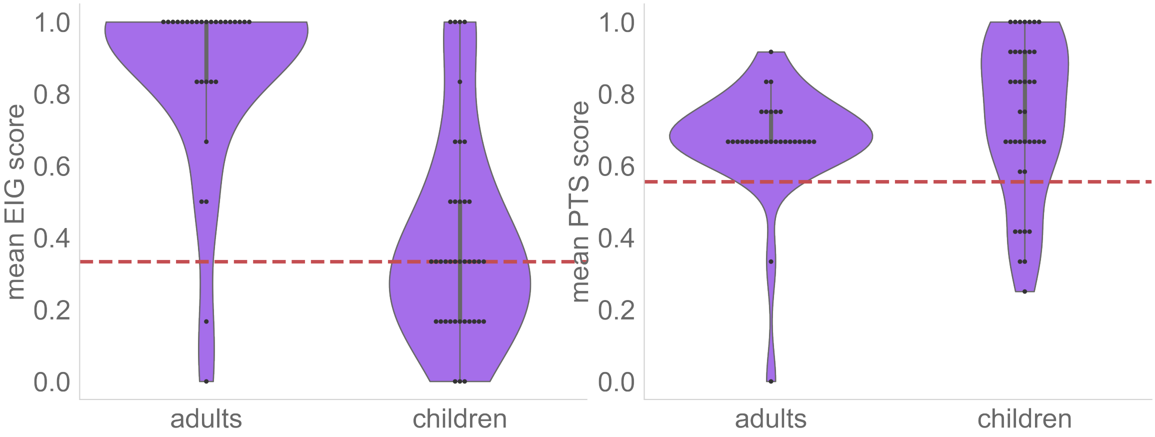
Participants

In Study 1, 39 5- to 7-year-old children ($M = 79.78$ months, $SD = 9.70$, range: 62–95 months) were tested at local schools and a children's museum and 29 adults ($M = 20.86$ years) at a public university. Before each child participated, the experimenter obtained informed consent from their legal guardian. Adults gave consent on their own behalf before participating for 0.5 course credit.

Procedure

In the very beginning, the experimenter demonstrated how to turn each light bulb on and off. In Study 1, the three light bulbs were controlled by three physical buttons of corresponding colors.

In the practice phase, the experimenter taught participants how to read graphs depicting four basic causal structures: 1) a one-link structure (🟡 → 🔴), 2) a causal chain (🟢 → 🟡 → 🔴), 3) a common-cause structure (🟡 ← 🟢 → 🔴), and 4) a common-effect structure (🟡 → 🔴 ← 🟢). When introducing the first example, the experimenter told children a cover story, *"Some light bulbs are special — they can light up other light bulbs! See the arrow going from the yellow to the red one? It tells us when the yellow turns on, it turns on the red* at the same time!" In subsequent examples, participants were told that the arrows had moved to different locations and asked to describe the new



(a) Mean EIG scores across participants and puzzles. **(b)** Mean PTS scores across participants and puzzles.

Figure 2.1: The mean EIG (left) and PTS (right) scores of the interventions chosen by participants in Study 1, averaged across all participants in each age group (adults *vs.* children) and all six puzzles they solved. Each dotted red line indicates the chance level of the respective score (EIG: 1/3; PTS: 0.55).

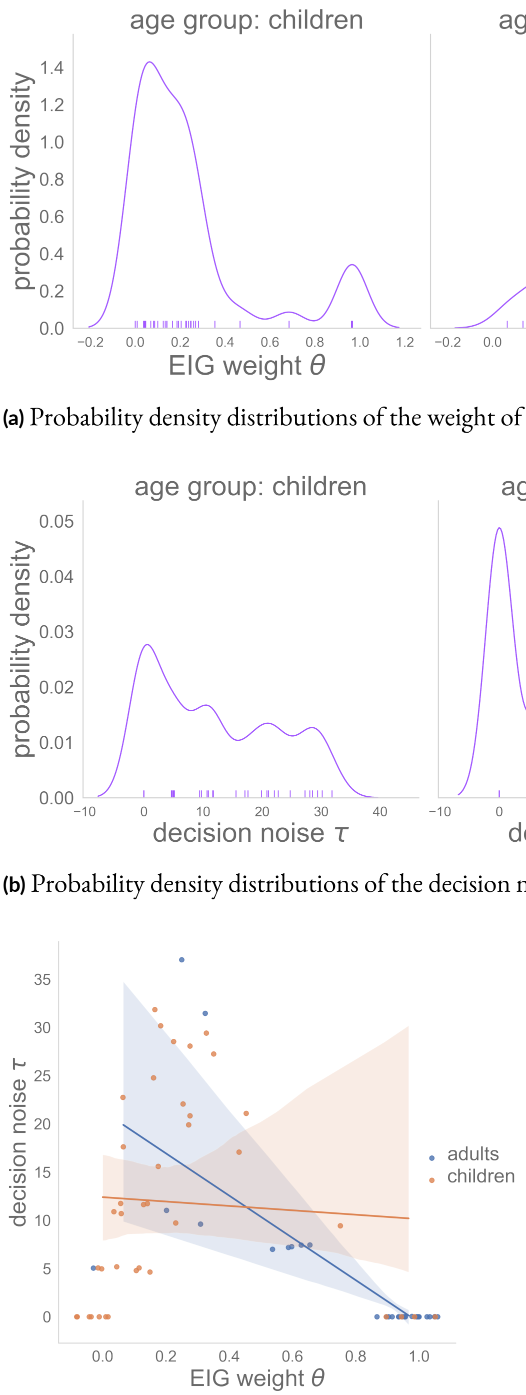
structure. After finishing describing each structure, participants were ask to predict what would happen when they turned on each light bulb. After making a prediction, participants turned on the said light bulb to see the effect for themselves. If participants made a wrong prediction, the experimenter would describe the effect, *"Um... See, the [color(s)] light bulb(s) turned on!"*

In the test phase, each participant solved six puzzles in a randomized order. In each puzzle, two possible structures were shown side by side. Participants were asked to describe each structure and then told that only one of them was correct about how the light bulbs actually work. The experimenter said, *"To find out which one is the answer, you can turn on one light bulb and see what happens. If the light bulb you turn on is useful, it can tell you the answer. However, not all light bulbs can help you — so choose carefully!"* After intervening on a light bulb and observing the outcome, participants were asked to choose the correct structure. No feedback was given during the study.

2.2 Results

Select interventions

Were participants able to select informative interventions? To answer this question, I first computed the mean EIG scores of children's and adults' chosen interventions across all six puzzles. If a learner chose interventions randomly, their expected EIG score would be 1/3 — in each puzzle, only one of the three light bulbs was informative. Adults' mean EIG score was .86 (95% CI [.76, .96]), which was far above chance, $t(28) = 10.95$, $p < .001$, Cohen's $d = 2.03$. Children's mean EIG score was .39 (95% CI [.30, .48]), which was at chance, $t(38) = 1.23$, $p = .23$, Cohen's $d = 0.20$.



(a) Probability density distributions of the weight of EIG θ in children and adults.

(b) Probability density distributions of the decision noise τ in children and adults.

(c) The weight of EIG θ as a function of decision noise τ.

Figure 2.2: Modeling results in Study 1: The weight of EIG τ (upper) and the decision noise τ (lower).

What did children do instead of maximizing EIG? Did they intervene randomly or use PTS? To shed light on alternative strategies, I computed each participant's mean PTS score across all puzzles. The chance level of PTS was .55, which was the mean PTS score across 18 interventions in all six puzzles. Children's mean PTS score was .74 (95% CI [.67, .81]), which was above chance, $t(38) = 5.37$, $p < .001$, Cohen's $d = 0.86$. Figure 2.1b shows each child's PTS score: As with the aggregated trend, most children's mean PTS scores were above the .55 chance level as well. These findings suggested that when children deviated from EIG, they were not intervening at random but following PTS instead. Adults' mean PTS score ($M = .67$, 95% CI [.61, .73]) was above chance as well, $t(28) = 3.95$, $p < .001$, Cohen's $d = 0.73$. This was a byproduct of the fact that the mean of PTS score of informative interventions (EIG = 1) was 2/3 and most adults chose informative interventions.

How can we formally characterize each participant's intervention strategy? To answer this question, I looked at each participant's weight of EIG θ (I provided a point estimate for each person using the mean of MCMC samples from their posterior distribution of θ). Across all adults, the mean of θ was .77 (95% CI [.65, .80]), suggesting that they used EIG more than half of the time, $t(28) = 4.74, p < .001$, Cohen's $d = 0.80$. Children's mean θ was .23 (95% CI [.20, .36]), suggesting that they used PTS more than half of the time, $t(38) = -5.75, p < .001$, Cohen's $d = 0.92$.

A learner who exclusively uses EIG or PTS has a decision noise τ close to 0; a learner who rarely chooses interventions with the highest EIG or PTS scores has a high decision noise. As we can see in Figure 2.2c, a roughly equal number of children were exclusive EIG and PTS users whereas when adults relied on one strategy, it was always EIG. In adults, τ decreased with θ, suggesting that those using EIG more were better able to maximize the values of chosen interventions. For children, the relationship between τ and θ was less clear. This may be because when adults strayed from EIG, it was mostly due to their inability to maximize intervention values; when children did so, it was either because they were unable to maximize intervention values or they purposefully used PTS.

Learn from interventions

Did participants learn from the outcomes of their own interventions? Adults did, but not children. Following informative interventions whose EIG = 1, adults' average accuracy was .92 (95% CI [.83, 1]), which was significantly higher than chance, $t(28) = 9.80, p < .001$, Cohen's $d = 1.85$. Even after choosing informative interventions, however, children's average accuracy was at chance ($M =$ .45, 95% CI [.31, .59]), $t(28) = -0.74, p = .46$, Cohen's $d = 0.12$. Individual data points in Figure 2.3 suggest that only a handful of children were actually choosing answers at random. Many children who chose incorrect structures did so consistently, indicating systematic misunderstandings of the link between intervention outcomes and the underlying structures that generated them.

2.3 Discussion

On the group level, adults mainly used EIG to select interventions whereas 5- to 7-year-olds mostly relied on PTS. On the individual level, most adults used EIG exclusively whereas most children mixed the two strategies. Among 39 children, about five solely relied on PTS and EIG, respectively.

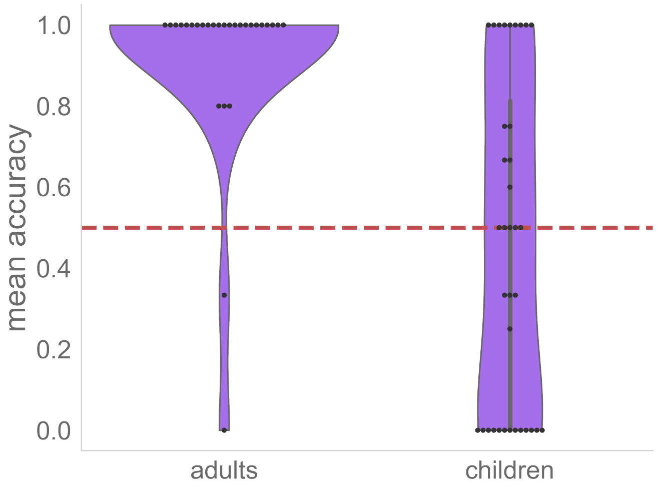


Figure 2.3: The average accuracy of identifying the correct structures among two after informative interventions in children (left) and adults (right) in Study 1. The dotted red line indicates the 1/2 chance level.

Even after informative interventions, fewer than half of the children performed above chance when identifying the correct structures. Together, results in Study 1 suggest that 5- to 7-year-olds were neither able to select informative interventions nor learn from the outcomes of their own interventions.

Taken at face value, findings in Study 1 seem at odds with a myriad of studies showing children's competence in causal learning (see Gopnik[53] for a review). Looking closer, the current study is different in several regards. To begin, children chose their own interventions in Study 1, as opposed to observing outcomes of interventions chosen by others. Moreover, the learning task was more challenging — instead of inferring the causal efficacy of one variable or the pairwise relationship between two variables, children in the Light Bulb Game needed to infer the structure of causal systems that had three nodes (representing variables) and three edges (representing causal relations). To our knowledge, only a handful of studies (McCormack et al.[83], Nussenbaum et al.[90]) had these components, including this dissertation. In such studies, children did *not* reliably outperform a random selection model until 8 years of age[83] or use EIG as their main strategy until around 12 years of age[90]. Maximizing EIG is genuinely difficult for young learners, especially in complex causal systems. Causality in the real world is rarely simple. An example still fresh in the public memory is that the chain of events leading up to the 2007 subprime mortgage crisis were so complex that they deceived the best of the economists. Facilitating children's intervention selection in complex causal systems is crucial to helping them skillfully navigate and make sense of the real world as adults.

Theories are good because they explain things, but explaining things turns out to be an awful lot like having theories of them.

Alison Gopnik (1998)

Chapter 3
Prompting to Explain

In the all-time classic courtroom drama *12 Angry Men*, a boy was heard yelling *"I'm going to kill you!"* at his father right before the latter dropped dead. Among the 12 jurors, 11 initially had no doubt that the boy was the culprit. *"Why would a murderer shout out a thing like that so the whole neighborhood could hear him?"*, Juror # 8 asked, raising a reasonable doubt in fellow jurors. In this case, explaining why someone did something helped the jury gauge the plausibility of the action.

Does explaining one's own action help them realize potential inefficiency and act more optimally in the future? In Study 1, 5- to 7-year-olds relied heavily on the suboptimal positive test strategy (PTS) to select interventions. Furthermore, they were at chance when identifying causal structures that led to the outcomes. Will asking learners to explain their intervention choices help them select better interventions and infer more accurately from the outcomes? To address this question in Study 2, I will prompt adults and children to explain why they plan to carry out a particular intervention and examine whether it impacts their intervention strategy and subsequent inferences.

The timing of the prompt is critical. When the explanandum is an observation (e.g., the ground is wet), we can ask learners to explain what they see (*"Why is the ground wet?"*), which may influence how they interpret the observation. The case of intervention selection is interesting: Asking for explanations after a choice has been made (*"Why did you turn on the green one?"*) cannot change the intervention choice retrospectively, yet before an intervention, there is no explanandum (turning on the green light bulb). To solve this problem, I will inform learners ahead of time that they will be asked to explain their intervention choices (*"After deciding on which light bulb to use, don't turn it on yet — point to it and I'll ask why you choose that one."*). After they point to a choice, I will prompt them again for an actual explanation, *"Why do you want to turn on that one?"*

If explaining does make a difference to learning, it may be attributed to act of explaining, or simply that the learners' attention was directed to the chosen interventions. In the self-explaining literature, a common control task is asking learners to report something instead of explaining it[132,133]. In Study 2, we can ask the control group to report their intended interventions, once before they make a choice (*"After deciding on which light bulb to use, don't turn it on yet — point to it and I'll ask which one you want to turn on."*) and again afterward (*"Which light bulb do you want to turn on?"*).

3.1 Method

Participants

In Study 2, 59 5- to 7-year-old children ($M = 81.86$ months, $SD = 9.31$, range: 61–101 months) and 30 adults ($M = 20.30$ years) participated in the Explanation condition. Another 58 5- to 7-year-olds ($M = 80.59$ months, $SD = 11.40$, range: 50–101 months) and 27 adults ($M = 20.44$ years) participated in the Report condition. Children were tested at local schools or a children's museum and we obtained informed consent form their legal guardians before the study. Adults were recruited and tested at a public university, each giving consent before participating for 0.5 course credit.

Procedure

The procedure in Study 2 was similar to that in Study 1, except that participants were prompted to explain or report their intervention choice in each puzzle. In both conditions, the experimenter alerted the participants, *"After deciding on which light to use, don't turn it on yet. Point to it first!"*

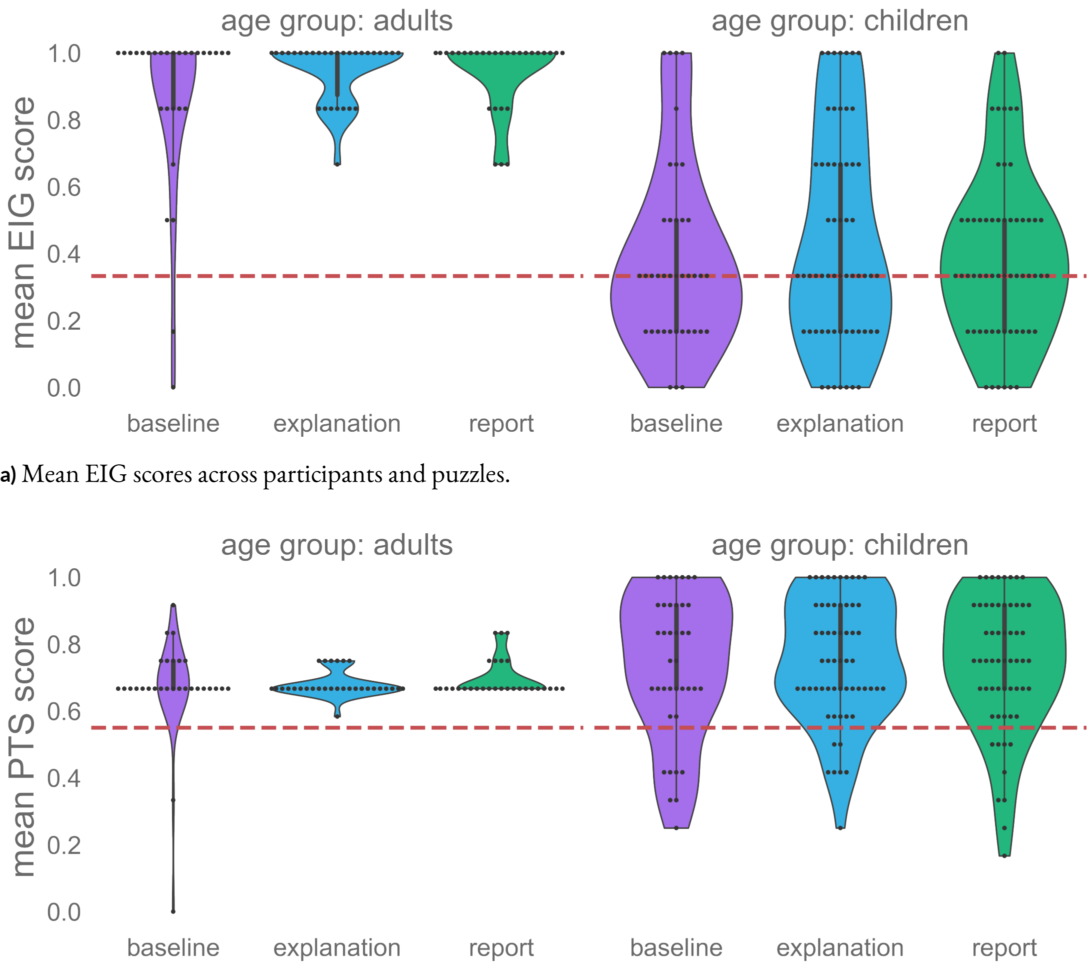
In the Explanation condition, the experimenter told participants, *"I'll ask you why you use that one to help."*, whereas in the Report condition, they said, *"I'll ask you which you wanna use to help."* After participants pointed to a light bulb, the experimenter asked them to either explain (*"Why do you choose that light bulb to find the answer?"*) or report (*"Which light bulb do you choose to find the answer?"*) their choice, depending on the condition. If participants changed their mind while explaining or reporting, their final choice was used in subsequent modeling and data analysis. No feedback on explanations or interventions was given during the study.

3.2 Results

Study 1 can be seen as the baseline of Study 2: The three conditions (Study 1: Baseline; Study 2: Explanation and Report) in the two studies shared similar procedures and the same age groups, except that participants in Study 1 solved puzzles on their own without being prompted to explain or report their intervention choices. For clearer comparisons, I pooled results from both studies together.

Select interventions

Perhaps surprisingly, explaining had no impact on learners' intervention strategies. Across all three conditions (Baseline in Study 1; Explanation and Report in Study 2), adults consistently chose informative interventions but not children. Adults' mean EIG scores were .95 (95% CI [.92, .98]) in the Explanation condition and .94 (95% CI [.90, .98]) in the Report conditions, both far above chance, $t(29) = 37.88, p < .001$, Cohen's $d = 6.92$, and $t(26) = 29.39, p < .001$, Cohen's $d = 5.46$.



(a) Mean EIG scores across participants and puzzles.

(b) Mean PTS scores across participants and puzzles.

Figure 3.1: The mean EIG (upper) and PTS (lower) scores of interventions chosen by participants in Studies 1-2, averaged across all puzzles solved by all participants in each age group (children *vs.* adults) and condition (Study 1: the Baseline condition; Study 2: the Explanation and the Report conditions). Each dotted red line indicates the chance level of the respective score (EIG: 1/3; PTS: 0.55).

Children's mean EIG scores were .43 (95% CI [.35, .52]) and .38 (95% CI [.32, .45]) in the Explanation and the Report conditions, respectively. The former score was above chance, $t(58) = 2.52, p = .014$, Cohen's $d = 0.33$, but not the latter, $t(57) = 1.63, p = .11$, Cohen's $d = 0.21$. Within each age group (children *vs.* adults), there were no differences in mean EIG scores between any conditions.

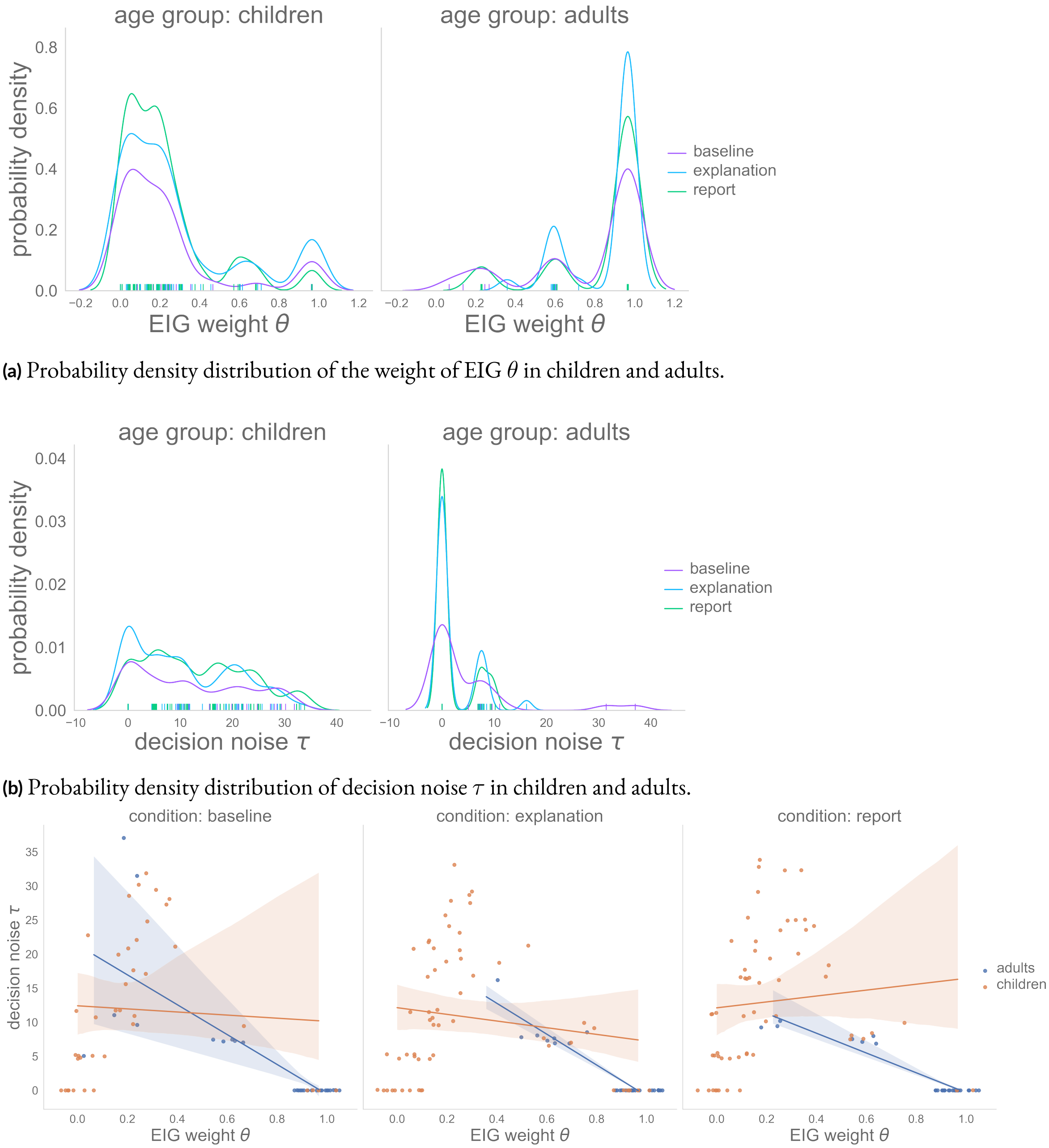
Modeling results suggested that adults mostly relied on EIG whereas children mainly used PTS. Adults' mean EIG weight θ was .86 (95% CI [.79, .93]) in the Explanation condition and .84 (95% CI [.75, .93]) in the Report condition, both above .50 (equal reliance on EIG and PTS), $t(29) = 10.97$, $p < .001$, Cohen's $d = 2.00$, and $t(28) = 7.41, p < .001$, Cohen's $d = 1.38$, respectively. Children's mean EIG weight θ was .28 (95% CI [.20, .36]) in the Explanation condition and .21 (95% CI [.16, .27]) in the Report condition, both less than .50, $t(58) = -5.42, p < .001$, Cohen's $d = 0.71$, and $t(57) = -9.79, p < .001$, Cohen's $d = 1.29$, respectively. Figure 3.2c shows that most children combined EIG and PTS to select interventions but several used PTS (τ and θ both close to 0) or EIG (τ close to 0 and θ close to 1) exclusively. When adults used a single strategy, it was always EIG.

Learn from interventions

In both the Explanation and the Report conditions in Study 2, adults and children identified the correct structures after informative interventions most of the time. In Study 1, only adults were able to do so. Adults' mean accuracy was .97 (95% CI [.93, 1]) in the Explanation condition and .94 (95% CI [.90, .99]) in the Report condition, both above the 1/2 chance, $t(29) = 25.18, p < .001$, Cohen's $d = 4.60$, and $t(28) = 19.87, p < .001$, Cohen's $d = 3.69$, respectively. Children's mean accuracy was .80 (95% CI [.72, .89]) in the Explanation condition and .65 (95% CI [.55, .76]) in the Report condition, also both above chance, $t(58) = 2.86, p < .001$, Cohen's $d = 0.99$, and $t(57) = 7.41$, $p < .001$, Cohen's $d = 0.40$, respectively. Between the two conditions in Study 2, children's mean accuracy in the Explanation condition was significantly higher than that in the Report condition, $t(102) = 2.15, p = .034$, Cohen's $d = 0.42$. Compared to those in the Baseline condition (mean accuracy: 44%), children in both the Explanation and the Report conditions were better able to learn from intervention outcomes, $t(96) = 3.23, p = .0016$, Cohen's $d = 1.01$, and $t(95) = 2.52$, $p = .013$, Cohen's $d = 0.52$, respectively.

3.3 Discussion

The biggest surprise in Study 2 is that prompting adults and 5- to 7-year-old children to explain their intervention choices did *not* change their intervention strategies: Across all conditions (Baseline in Study 1; Explanation and Report in Study 2), adults mainly used EIG and children PTS. The only notable difference between the two studies was that children in both the Explanation and the Report conditions in Study 2 chose the correct structures more than half of the time, which was an improvement from the chance-level accuracy in Study 1. Explainers improved more than those in the control group who only reported their choices. A possible reason behind the improvement in accuracy is that directing children's attention to their intervention choices made them more aware of how the intervention outcomes related to the underlying causal structures that generated them.



(a) Probability density distribution of the weight of EIG θ in children and adults.

(b) Probability density distribution of decision noise τ in children and adults.

(c) The weight of EIG θ as a function of decision noise τ.

Figure 3.2: Modeling results in Studies 1-2: The weight of EIG τ (a) and decision noise τ (b).

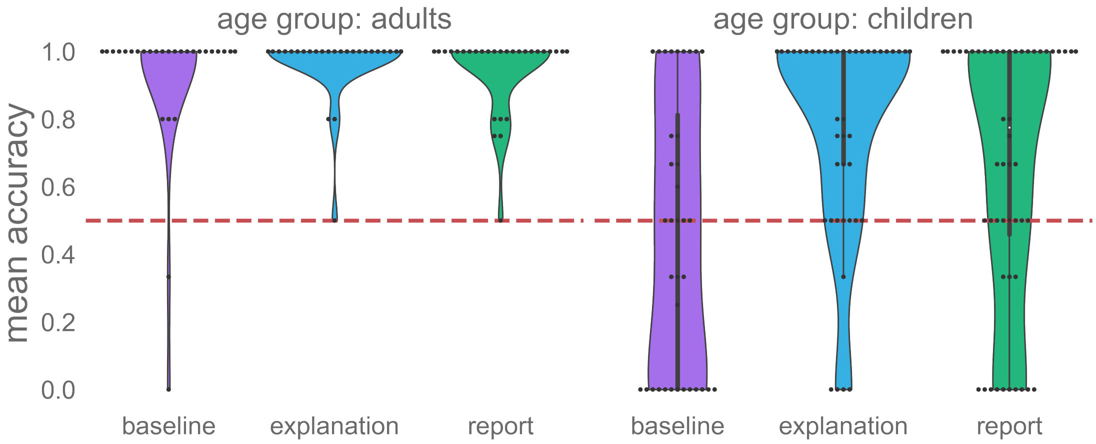


Figure 3.3: The average accuracy of identifying the correct structures among two after informative interventions in children (left) and adults (right) in Studies 1-2. The dotted red line indicates the 1/2 chance level.

The lack of self-explaining effects on intervention selection was somewhat surprising. Typically in self-explaining studies, engaging in the act of explaining can already facilitate learning, even when learners don't obtain the correct explanations[22,77,135]. How does the current study differ from the past? I argue that a major difference lies in the nature of the explanandum. In almost all past studies, the explanandum is an observation, such as the category membership of an exemplar or the causal efficacy of an object. Take the now classic work by Williams and Lombrozo for instance[136]: Even when someone fails to explain that an alien is a Glorp because it has pointy feet, they are still attending to other visual features of the exemplar, such as its color or body shape. Incorrect explanations still pertain to the categorization task and may therefore impact the categorization results. In the Light Bulb Game, however, the explanandum is an intervention, which is harder to explain.

How should you explain why you plan to turn on a certain light bulb? First of all, you must recognize your own uncertainty, as opposed to assuming you already know the true structure before you intervene (like some children did). Then, you must enumerate all possible interventions, simulate all possible outcomes following each intervention, and re-assess to what degree each outcome can reduce your uncertainty. Most 5- to 7-year-olds' explanations (e.g., *I chose the red one because red is my favorite color.*) did not pertain to any of the steps involved in intervention selection.

I hypothesize that, in order for explaining to facilitate intervention, learners need to generate explanations that target the intervention selection process in some way. For instance:

- Awareness of uncertainty: *"Because X can help me be sure which picture is correct."*
- Outcome simulation: *"If I turn on X, A happens in the left picture and B in the other."*
- Outcome comparison: *"A and B are different outcomes."*

- Uncertainty reassessment: *"If A happens, I'm sure the left picture is correct; if B happens, on the other hand, I'm sure the right one is correct."*

To test the above hypothesis, we can study older children who are presumably better explainers to see if they also intervene better. It's also more practical to test older children remotely* and get intelligible explanations. However, age-related variables such as intelligence and formal education may affect both how learners select interventions and how they explain intervention choices, making it hard to claim a causal connection between the two. To make Study 3 more practical without introducing the confounders, I will test 7- to 11-year-olds and manipulate their explanation quality by training them on EIG maximization and examine whether they rely more on EIG as a result.

*As I was designing Study 3, the COVID pandemic broke out and in-person testing was moved to Zoom.

It is the usual fate of mankind to get things done in some boggling way first, and find out afterward how they could have been done much more easily and perfectly.

Charles S. Pierces (1882)

Chapter 4
Training to Explain

In Study 2, asking children to explain their intervention choices didn't impact their intervention strategies: They still mostly relied on PTS rather than the optimal strategy, EIG maximization. As discussed at the end of Chapter 3, it could be that interventions are so hard to explain that merely engaging in explaining does not impact intervention selection. To provide children with more effective scaffolding, in Study 3, I will train them to explain interventions using the *difference-making* principle (e.g., *"I choose the [color] light bulb because it turns on different light bulbs in each picture."*), which is at the core of EIG maximization, and examine whether they choose more informative interventions later on. During training, children are asked to explain whether or not each intervention choice is useful, *"Can the [color] light bulb help you find the answer? Why (not)?"*. If a child's explanation doesn't follow the difference-making principle, the experimenter will directly teach this principle to them (see Procedure for more details). To measure potential training effects, I will test another group of children in the Explanation condition (identical to that in Study 2), where they are prompted to explain their intervention choices without being trained to do so.

If EIG training does make a difference, it could be because children were asked to explicitly explain whether each intervention was usefulness, or that training forced them to pay attention to all the possible interventions, not just the ones they selected. It's an interesting empirical question as to whether explicit explanations of usefulness are necessary to influence explanations and interventions, or if attending to all intervention choices suffices. To answer this question, I will ask another group of learners to describe what each intervention does in each picture, *"Can you tell me what the [color] light bulb does in the left/right picture?"*, without explaining whether it makes the said intervention helpful. When designing Study 3, preliminary results showed that only 9- to 11-year-olds benefited significantly from EIG training. As a result, I only tested this age group in the Control condition*.

*After analyzing Study 3 data, however, I found that 7- to 8-year-olds also benefited from training, albeit to a lesser degree than the older children. In a follow-up study, I will test the younger children in the Control condition. The new condition was not proposed as part of the dissertation and will be published afterward.

4.1 Method

Participants

In Study 3, 61 7- to 11-year-olds ($M = 106.60$ months, $SD = 13.39$, range: 85–131 months) participated in the Explanation condition and another 57 ($M = 106.58$ months, $SD = 15.53$, range: 84–140 months) in the Training condition. I used the median age of the 118 children (108 months) to split each condition into two age groups, resulting in 31 7- to 8-year-olds ($M = 95.28$ months, $SD = 6.96$, range: 85–107 months) and 30 9- to 11-year-olds ($M = 118.30$ months, $SD = 6.49$, range: 108–131 months) in the Explanation condition and 27 7- to 8-year-olds ($M = 93.18$ months, $SD = 8.07$, range: 84–107 months) and 30 9- to 11-year-olds ($M = 118.63$ months, $SD = 9.51$, range: 108–140 months) in the Training condition. Another 24 9- to 11-year-olds ($M = 120.62$ months, $SD = 7.48$, range: 109–140 months) were tested in the Control condition. Before each child participated over Zoom, we obtained informed consent from their legal guardians via Qualtrics.

Procedure

The Explanation condition in Study 3 was identical to that in Study 2. All but the first trial in the Training and the Control conditions in Study 3 were the same as that in the Explanation condition.

In the Training condition, whichever puzzle that appeared first (e.g., Figure 4.1) was used to train participants on EIG maximization. The experimenter randomly picked a light bulb and asked, *"If we turn on the [color] light bulb, can it help our friend Alex find out the answer?"* If participants answered correctly, the experimenter moved on to the next light bulb and asked the same question. If participants gave incorrect responses, the experimenter asked them to describe what the said light bulb does in each structure, *"Let's think again. If the left picture was the answer, what would the [color] light bulb do? If the right picture was the answer, what would it do?"* After participants answered, the experimenter asked again, *"So, can this light bulb help Alex find out which of the two pictures is the answer?"* If participants answered correctly, the experimenter moved on to the next light bulb. If not, the experimenter taught participants the difference-making principle,

- If the choice was in fact uninformative: *"Um, I think the [color] one doesn't help, because it turns on the same light bulbs in both pictures. When you see [outcome], both small pictures might be correct, so you cannot tell which one matches the big picture."*
- Or if it was actually informative, *"Um, I think it helps — if A happens, you know the first picture is correct; if B happens instead, you know the second picture is correct."*

After participants correctly analyzed the usefulness of each intervention choice, the experimenter told them to choose one intervention to solve the puzzle. If the chosen light bulb was uninformative, participants were asked to re-describe what it does in each picture, *"Um, let's think again. What does the [color] light bulb do in the left picture? What does it do in the right picture?"*, and

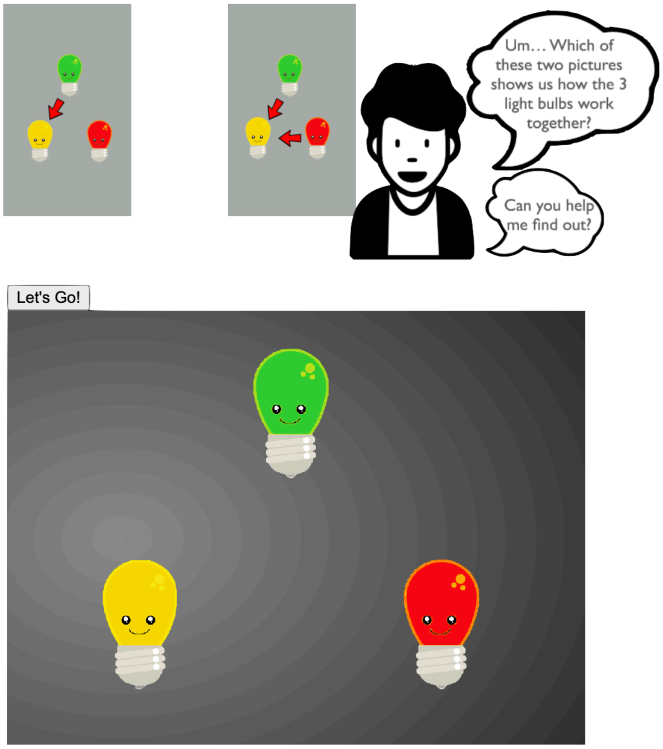

Um… Which of these two pictures shows us how the 3 light bulbs work together?
Can you help me find out?
Let's Go!

Figure 4.1: In the Training condition, the first puzzle was used for EIG training: Participants were asked to explain whether or not each light bulb could solve the puzzle. In the Control condition, participants were asked to report what each light bulb does in the two structures, without explaining its informativeness.

re-think its usefulness, *"So if [outcome] happens, can you tell if the left or the right picture is the answer?"* If participants didn't answer correctly, the experimenter provided them with EIG-based reasoning again, *"I don't think it helps because both pictures could be correct when [outcome] happens!"*, or, *"I think it's useful, because if A happens, we know it's the left picture, but if B happens, we know it's the right one!"* After choosing and turning on an informative light bulb, participants were asked to identify the correct structure, for which they received feedback, *"That's (not) the right answer!"*.

The Control condition was similar to the Training condition. Instead of explaining whether each light bulb could help solve the puzzle, participants were asked to describe its effects in both structures, *"If we turn on the [color] light bulb, what does it do in the two pictures?"* If participants described correctly, the experimenter moved on to the next light bulb. If participants were incorrect, the experimenter asked them to describe again, *"Let's think again. In the left picture, what does the [color] light bulb do? In the right picture, what does it do?"* The remaining procedure was again similar to that in the Training condition, except when participants chose an uninformative light bulb, the experimenter simply asked them to pick a different one until correct, *"Let me check my notes. Actually, the [color] light bulb cannot help. Which other light bulb do you wanna use?"*

explanation type	example
EIG	“green only turns on red in one of two pictures”
PTS	“yellow turns on all three no matter what”
describe one	re-describe one of the pictures
describe both	re-describe both pictures
goal only	“because it helps me find the answer”
prior belief	“because the picture on the right is the answer”
guess	“I don’t know; just feel like it”
other	“because I like the color red”
no explanation	child didn’t explain or experimenter didn’t ask

Table 4.1: Nine types of explanations generated by children in Study 3 and a typical example of each type.

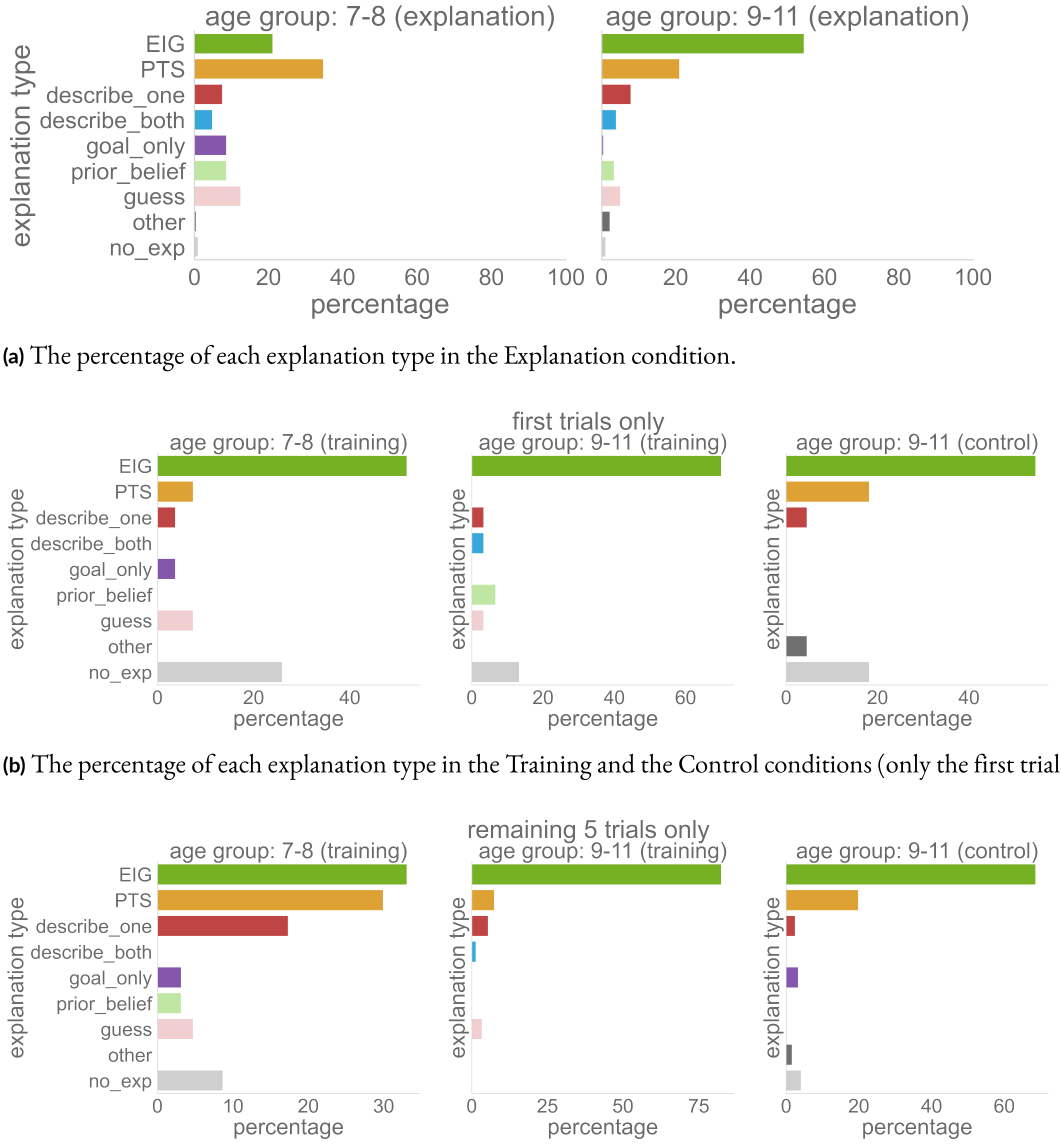
4.2 Results

Explanations types

I coded the explanations generated by children into 9 types (Table 4.1). Explanations were coded as EIG if they mentioned an intervention was useful because it could lead to different outcomes in different structures or that it was unhelpful because it would make the same thing happen in both structures. Explanations were coded as PTS when they either mentioned an intervention could potentially activate all light bulbs, or that it would activate the same light bulbs in both structures. If a child described what all three light bulbs would do in one or both pictures, such explanations were coded as “describe one” or “describe both”. Sometimes children only said they chose an intervention because it would help them solve the puzzle, without explaining how it could do that. Such explanations were coded as “goal only”. If a child said they didn’t know why they chose an intervention, their explanation was coded as “guess”. In some cases, children provided irrelevant explanations or didn’t explain — the former was coded as “other” and the latter “no explanation”.

In the Explanation condition (Figure 4.2a), untrained 9- to 11-year-olds already showed a greater tendency to explain interventions in terms of EIG than 7- to 8-year-olds. Among 7- to 8-year-olds, PTS was the most popular type of explanation, occurring 34.78% of the time, and EIG was the second most popular type, occurring 21.20% of the time. Among 9- to 11-year-olds, EIG was the most popular type, occurring 54.54% of the time, followed by PTS, which occurred 21.20% of the time.

Did EIG training lead children to generate more EIG-based explanations? It did. In the Training condition, on the first trial which was used for EIG training (Figure 4.2b), 7- to 8-year-olds generated EIG-based explanations 51.85% of the time, which was about 30% more frequently than those in the Explanation condition; 9- to 11-year-olds generated EIG-based explanations 70% of the time, which was an 15% increase from the Explanation condition. For 9- to 11-year-olds, those in the Control condition generated EIG-base explanations just as frequently (54.54%) as those in the Explana-



(a) The percentage of each explanation type in the Explanation condition.

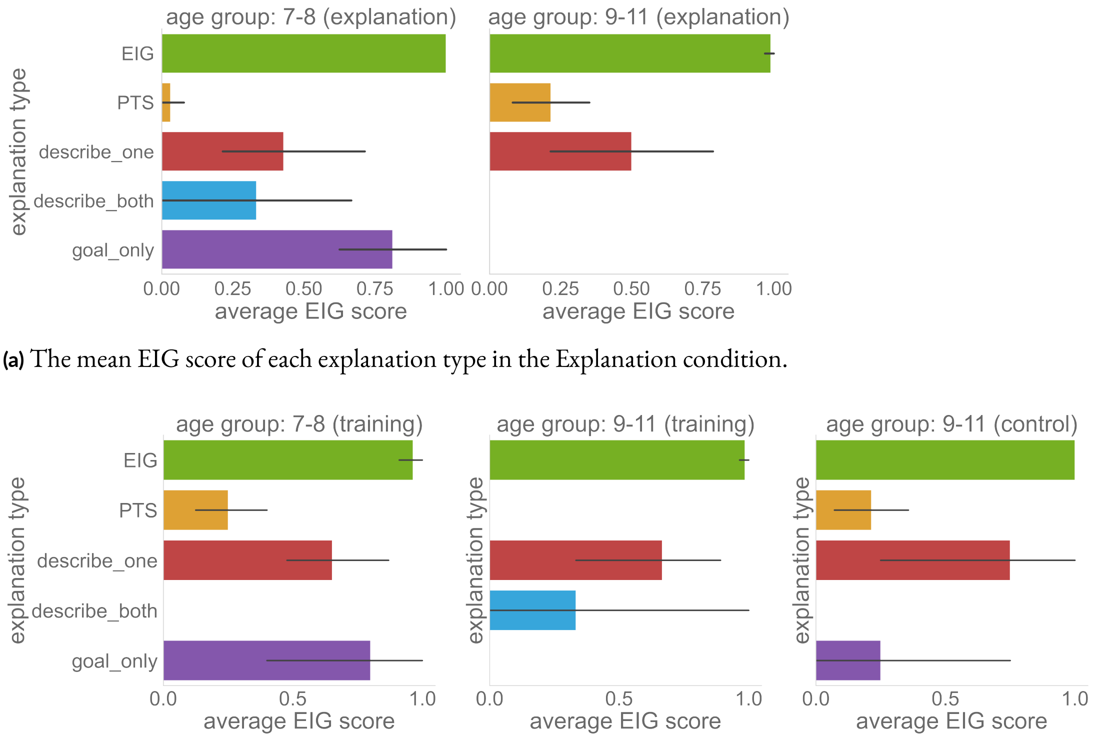
(b) The percentage of each explanation type in the Training and the Control conditions (only the first trials).

(c) The percentage of each type of explanation in the Training and the Control conditions (remaining 5 trials).

Figure 4.2: The percentage of each type of explanation by 7- to 8-year-olds (left) and 9- to 11-year-olds (right).

tion condition. In the Training condition, the training effect on the explanation type lasted longer in older children than the younger ones. During the remaining 5 trials, 7- to 8-year-olds generated EIG-based explanations 33.07% of the time, whereas 9- to 11-year-olds did so 82.42% of the time (Figure 4.2c). In the Control condition, 9- to 11-year-olds provided EIG-based explanations 68.59% of the time in remaining trials, which was in between the Training and the Explanation conditions.

Select interventions



(a) The mean EIG score of each explanation type in the Explanation condition.

(b) The mean EIG score of each explanation type in the Training and the Control conditions.

Figure 4.3: The mean EIG score of each explanation type in Study 3.

A learner may know that, *in principle*, they should intervene on a light bulb that leads to distinct outcomes in different structures, yet *in reality*, cannot find which one it is. To see if interventions participants chose matched explanations that they gave earlier, I plotted the mean EIG scores for the five most common types of explanations. In both age groups and all conditions (Figure 4.3), children who provided EIG-based explanations almost always chose informative interventions; those who gave PTS-based explanations rarely did. Patterns of other types were less clear. Notably, 7- to 8-year-olds who provided "goal only" explanations chose informative interventions most of the time, suggesting they might think of EIG as the reason why the chosen intervention could help.

Overall, did training improve intervention selection? It did, especially for 9- to 11-year-olds in the Training condition. Seven- to 8-year-olds' mean EIG score in the Training condition was .52 (95% CI [.39, .66]), which was significantly higher than their mean EIG score in the Explanation condition ($M = .42$, 95% CI [.29, .55]), $t(54.99) = 2.13, p = .038$, Cohen's $d = 0.56$. Nine- to 11-year-olds' mean EIG score in the Training condition was .84 (95% CI [.73, .95]), which was also significantly higher than their mean EIG score in the Explanation condition ($M = .64$, 95% CI [.50, .77]), $t(54.99) = 2.90, p = .005$, Cohen's $d = 0.75$. However, older children's mean EIG score in the Control condition ($M = .75$, 95% CI [.59, .91]) was not significantly higher than that in the Explanation condition, $t(50.93) = 1.50, p = .14$, Cohen's $d = 0.41$. The mean EIG weight θ of 7- to 8-year-olds was .26 (95% CI [.13, .39]) in the Explanation condition and .35 (95% CI [.20, .50]) in the Training condition; the increase in the Training condition was not statistically significant, $t(46.66) = 0.93, p = .36$, Cohen's $d = 0.26$. Nine- to 11-year-olds' mean EIG weight θ was .72 (95% CI [.59, .85]) in the Training condition, which was significantly higher than that in the Explanation condition ($M = .45$, 95% CI [.30, .60]), $t(56.14) = 2.76, p = .008$, Cohen's $d = 0.72$. Notably, the performance of 9- to 11-year-olds in the Training condition was on a par with adults in Study 1's Baseline condition (mean EIG score = .86; mean EIG weight $\theta = .77$). Again, in the Control condition, 9- to 11-year-olds mean EIG weight θ ($M = .63$, 95% CI [.46, .80]) did not statistically differ from that in the Explanation condition, $t(49.17) = 1.63, p = .11$, Cohen's $d = 0.45$.

θ increased with age in the Explanation and the Training conditions (Figure 4.6). Across all ages, children in the Training condition tended to have higher θ than those in the Explanation condition[†].

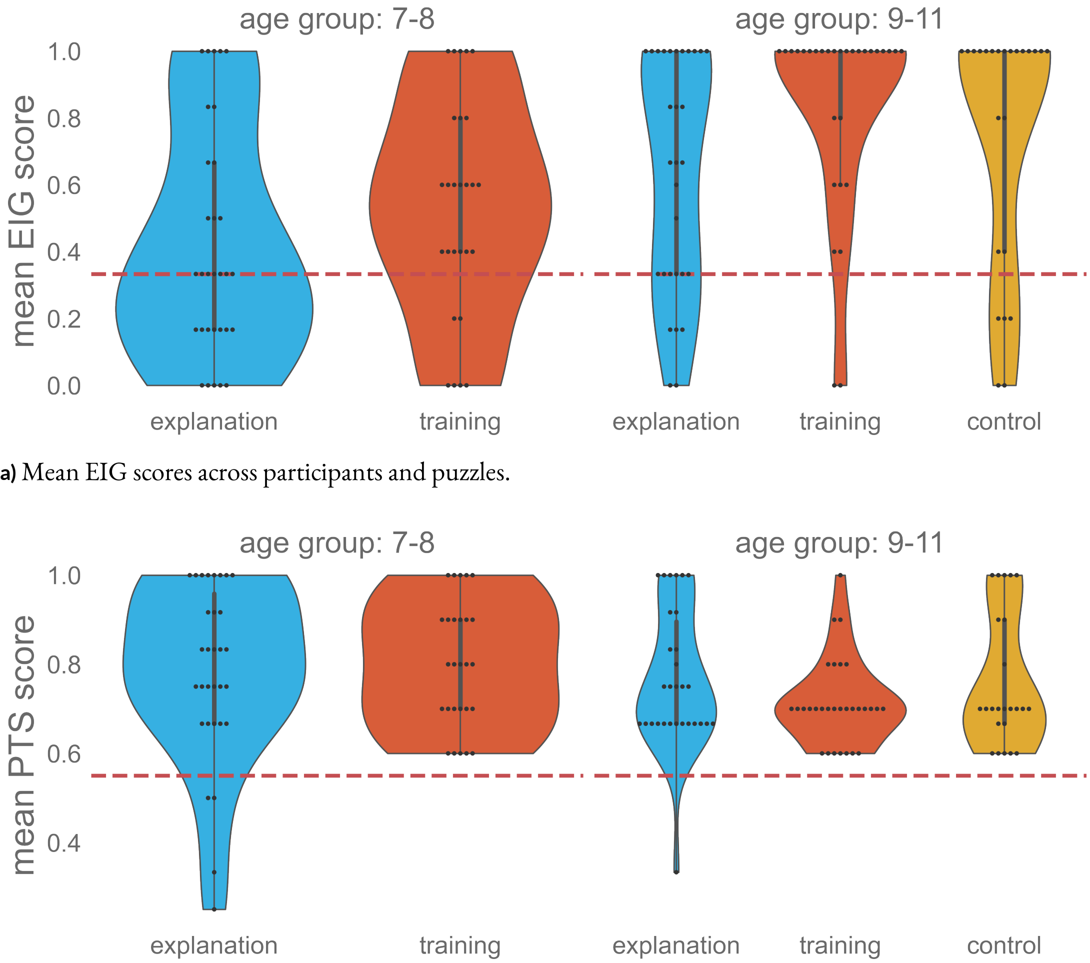
Learn from interventions

Across conditions and age groups, children were able to identify correct structures after informative interventions most of the time. Seven to 9-year-olds' mean accuracy was .90 (95% CI [.78, 1]) in the Explanation condition and .87 (95% CI [.74, .99]) in the Training condition, both above the 1/2 chance, $t(25) = 7.15, p < .001$, Cohen's d = 1.40, and $t(21) = 6.00, p < .001$, Cohen's d = 1.28, respectively. Nine- to 11-year-olds' mean accuracy was .94 (95% CI [.88, 1]) in the Explanation condition, .96 (95% CI [.91, 1]) in the Training condition, and . 98 (95% CI [.93, 1]) in the Control condition, all above chance as well, $t(27) = 13.94, p < .001$, Cohen's d = 2.63, $t(28) = 17.46$, $p < .001$, Cohen's d = 3.24, and $t(21) = 21.00, p < .001$, Cohen's d = 4.47, respectively.

4.3 Discussion

In Study 3, briefly training 7- to 11-year-olds on EIG maximization for a single trial led them to generate a higher percentage of EIG-based explanations for their intervention choices as well as relying more on EIG to select interventions. The training effect was more pronounced in 9- to 11-year-olds than it was in 7- to 8-year-olds. Regardless of the age group and the condition, children were able to identify the correct structures from the outcomes of their own interventions.

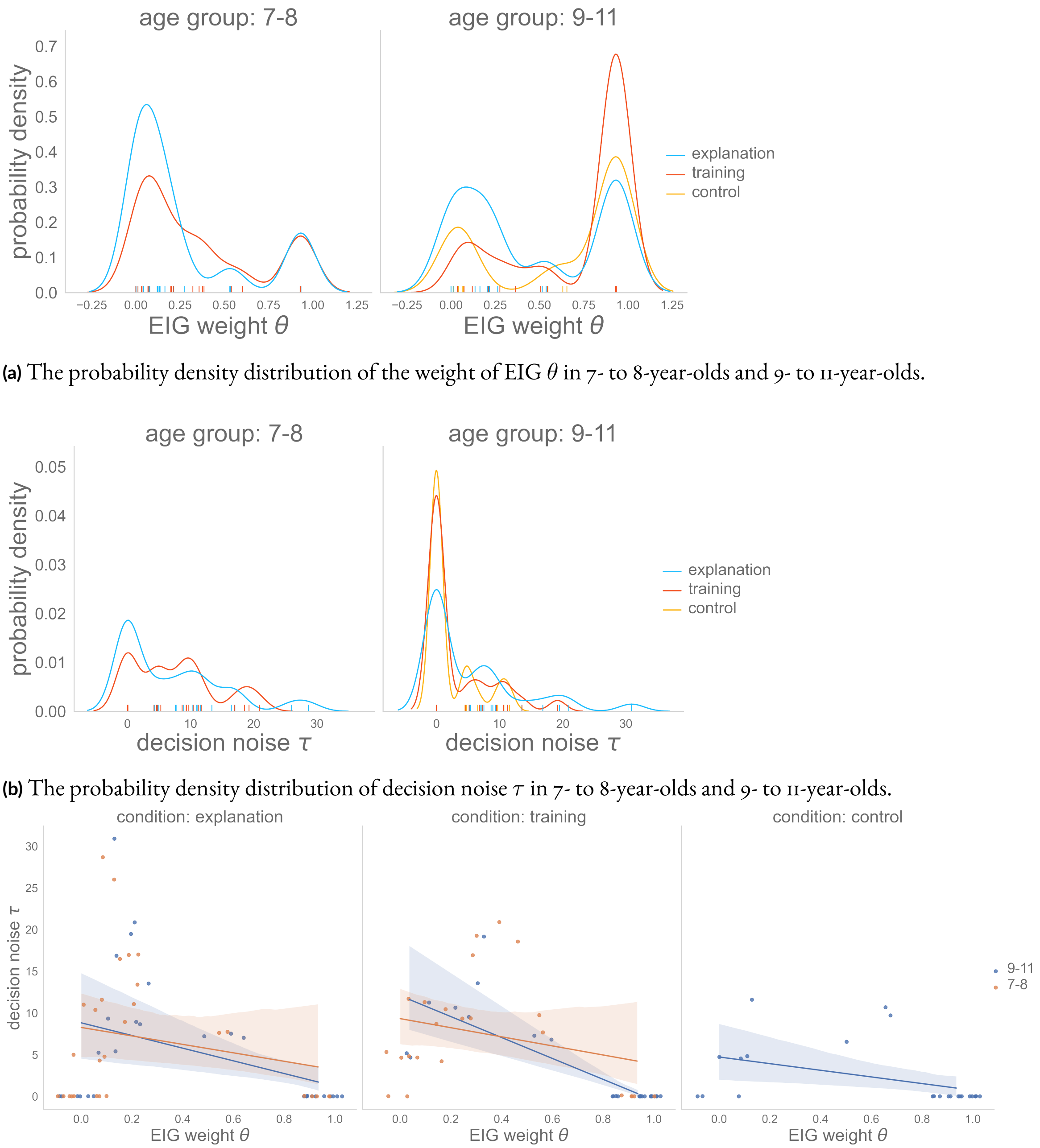
[†] In the Control condition, θ decrease with age, which may well change with data from 7- to 8-year-olds.



(a) Mean EIG scores across participants and puzzles.

(b) Mean PTS scores across participants and puzzles.

Figure 4.4: The mean EIG (a) and PTS (b) scores of interventions chosen by participants in Study 3, averaged across all non-training puzzles solved by participants in each age group (7- to 8-year-olds *vs.* 9- to 11-year-olds) and condition. Each dotted red line indicates the chance level of the respective score (EIG: 1/3; PTS: 0.55).



(a) The probability density distribution of the weight of EIG θ in 7- to 8-year-olds and 9- to 11-year-olds.

(b) The probability density distribution of decision noise τ in 7- to 8-year-olds and 9- to 11-year-olds.

(c) The weight of EIG θ as a function of decision noise τ.

Figure 4.5: Modeling results in Study 3: The weight of EIG τ (upper) and decision noise τ (lower).

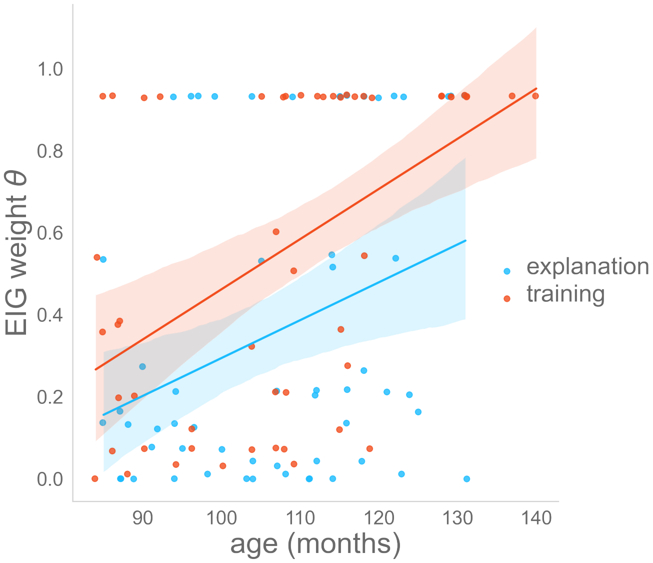


Figure 4.6: Weight of EIG θ as a function of children's age and the condition (Explaining *vs.* Training).

How did EIG training impact children's intervention selection? One possibility is that, the training effect comes from children explaining the usefulness of each intervention choice. Another is that training forced children to pay attention to all intervention choices before settling on one, whereas in the Explanation condition, they were not required to think it through (e.g., a child may say *"I want to turn on the yellow one"* without considering other light bulbs). The former seemed more likely given the results in Study 3: In the Control condition where children were only asked to describe what each light bulb does in each picture, their intervention strategies did *not* differ significantly from that in the Explanation condition — it was only in the Training condition where children explicitly explained the usefulness of each choice that we found significant improvements.

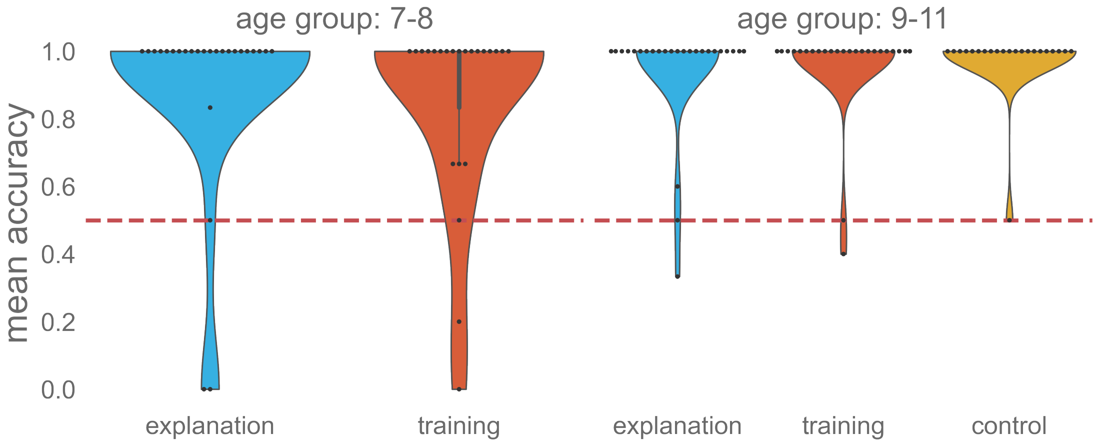


Figure 4.7: The average accuracy of identifying the correct structures after informative interventions in 7- to 8-year-olds (left) and 9- to 11-year-olds (right) in Study 3. The dotted red line indicates the 1/2 chance level.

Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education, one would obtain the adult brain.

Alan Turing (1950)

Chapter 5
Conclusion

In merely 10 years, machines went from barely able to recognize objects[34,64] to competently playing video games, holding conversations, captioning images, and controlling robot arms using a unified model trained on a wide range of data[104]. Are large neural nets the end-all and be-all of intelligence? Anyone who has observed a child would say no: Parents don't raise a child by following them around and feeding them billions of examples; rather, the child actively *inquires* about the world through experimentation, questioning, and exploration (learning by experimentation) and *thinks* about existing data to draw generalizable conclusions (learning by thinking). In this dissertation, I explored the boundaries of learning by thinking and discovered a new way to facilitate experimentation: Briefly training 7- to 11-year-olds to understand good experiments lead to different outcomes under different hypotheses helped them design informative experiments on their own.

To illustrate key ideas and findings from this dissertation, I invite you to an ordinary evening.

5.1 Summary — *Fiat Lux*

I was writing this dissertation when the lights in my room went out. I stood up to check my neighbors' windows. Their lights were on. I went to open my fuse box, saw the bedroom trip switch was in the "off" position, and pushed it "on". The lights turned on and I went back to writing.

In a split second, I observed a sudden change in my room, generated multiple hypotheses (*Was there a PG&E outage? Did the circuit breaker trip? Did all light bulbs burn out?*), and assessed their prior probabilities (the first two seemed likely but it would be wicked if all light bulbs burned out at once). After that, I performed an informative intervention (checking my neighbors' lights) and learned from the outcome (had there been an outage, my neighbors would be in the dark, too — instead, something likely happened to me alone). The newly gained causal knowledge served my immediate and long-term goals: Getting the lights back on and getting my dissertation done.

This "superpower" is anything but unique to me — using interventions to shed light on causal structures is a hallmark of human intelligence[66,96]. Finding good interventions isn't always easy. As a child, my knee jerk reaction was often to turn the light switches off and back on. Since this action results in the same effect (or lack thereof) under hypotheses listed above, it leaves me "in the dark".

If I traveled back in time and asked my 7-year-old self, *"Hey, why do you wanna fiddle with the light switches?"*, would they realize the original plan was silly and do what I'd do now? Unfortunately, the little me just shrugged, *"No idea. Just felt like it."*, and repeated the same mistake. Wanting to help more, I invited the 7-year-old to think through plausible causes (outage, tripped breaker, burnout), enumerate potential interventions (checking on neighbors, looking into the fuse box, or changing bulbs), and imagine how much each intervention could narrow down the hypothesis space. The child now realized, whatever they do, it should make a difference in different cases.

In three studies reported in this dissertation, I have investigated the questions illustrated above: Without any help, can causal learners choose informative interventions and learn from intervention outcomes (Study 1 in Chapter 2)? Does prompting learners to explain their intervention choices help them choose better interventions (Study 2 in Chapter 3)? Does training learners to discriminate between hypotheses help them both better explain and select interventions (Study 3 in Chapter 4)?

In Study 1, I tested adults and 5- to 7-year-olds in the Light Bulb Game, where they intervened on one light bulb to identify the true structure of three light bulbs from two hypotheses. Each participant solved 6 puzzles. Based on their choices, I used a Bayesian model to capture how much each person relied on the optimal strategy, which is maximizing the chosen intervention's expected information gain (EIG), and the suboptimal positive test strategy (PTS). Adults mainly used EIG to select interventions and learned from outcomes of their own interventions. By contrast, children mainly used PTS and couldn't identify the correct structure even after an informative intervention.

In Study 2, I took a common approach in the self-explaining literature (a subdomain of learning by thinking), asking adults and 5- to 7-year-olds to *explain* why they chose certain interventions (the Explanation condition: *"Before you turn on a light bulb, point to it and I'll ask you why you wanna use it to help." "Can you tell you why you use the yellow light bulb to help?"*) to see if they might intervene differently. In the Report condition, additional groups of adults and children were asked to *report* their choices (*"Before you turn on a light bulb, point to it and I'll ask you which one wanna use to help." "Can you tell you which light bulb you wanna use to help?"*). Explaining didn't change learners' intervention strategies: In both the Explanation and the Report conditions, adults mainly used EIG and children PTS. However, children in Study 2 chose the correct structures after informative interventions most of the time, suggesting that *attending to* (explaining or reporting) one's intervention choices might help them connect intervention outcomes to underlying hypotheses.

In Study 3, I trained 7- to 11-year-olds on EIG maximization: In the first puzzle used as a training example, children were asked to explain whether or not each of the three light bulbs was useful. If they didn't provide EIG-based explanations themselves, the experimenter taught children the correct explanation (*"I think [color] light bulb was useful: If A happens, we know it's the left picture; if B happens, we know it's the right picture" "I think [color] light bulb was not useful: In both pictures, C would happen, so we can't tell which picture is correct!"*). Besides the Training condition, I tested another two groups of children. One group participated in the Explanation identical to that in Study. The other group participated in the Control condition: Instead of explaining whether each intervention was useful, they simply reported what it would do under the first and the second hypotheses, respectively. I used a median split to divide children into two age groups: 7- to 8-year-olds and 9- to 11-year-olds. Compared to their counterparts in the Explanation, both the younger and the older children generated a higher percentage of EIG-based explanations in puzzles they later solved on their own and relied more heavily on EIG to chose interventions; the effects were stronger in the older children compared to the younger children. Children in the Control condition saw some improvements but were not significantly different from those in the Explanation condition.

In summary, explaining alone doesn't lead to better interventions, but training one to explain might, especially older children who already use the optimal strategy more than the younger ones.

5.2 Contributions

This dissertation advances the field of learning by thinking by discovering the limit of merely explaining and learning by experimentation by finding a new way to facilitate intervention selection.

5.2.1 Boundary condition of self-explaining

In scientific education[21,22,23,43], human learning[77,135], and more recently, machine learning[67], self-explaining is wielded as a simple yet power tool to support learning and generalization, in a variety of ways across a variety of domains. For instance, compared to those not asked to explain, explainers prefer simple causal hypotheses that have the same scope as more complex ones[129], discover broad rules that categorize all exemplars rather than obvious yet inconclusive rules[136], and privilege abstract, generalizable knowledge over the concrete or the superficial[131,132], *etc.*. Even more conveniently, explanations don't have to be correct for explainers to benefit from explaining[21,22,23,43,77,135].

However, self-explaining is not a panacea for all learning obstacles. In Study 2, I found that 5- to 7-year-olds who were merely asked to explain their intervention choices (*"Why do you wanna use the yellow light bulb to help?"*) used similar strategies as those who tackled the same problems without explaining. I hypothesized that in order for explaining to facilitate learning, it needs to tap into the learning process. When learning category memberships of exemplars, incorrect explanations may refer to incorrect features (e.g., color but not shape, if shape is the correct rule), but attending to incorrect features nevertheless invokes the categorization process. When learning the causal efficacy of an object (e.g., whether it makes a machine play music), incorrect explanations may interpret contingencies between events incorrectly (e.g., forgetting the object never made the machine go in the absence of another one), but thinking about contingencies between events is still relevant to causal inference. By contrast, when explaining intervention choices, most 5- to 7-year-olds didn't shed light on the intervention selection process (i.e., being aware of own uncertainty, enumerating hypotheses and interventions, reassessing uncertainty after an intervention and all potential outcomes, etc.) at all, citing irrelevant reasons instead (e.g., *"I like the red light bulb"*, *"I just wanna try this"*).

When task-relevant explanations are difficult to obtain, self-explaining may have limited impact on learning — training learners to explain in the right way may help them benefit more from explaining. In Study 3, I tested children aged between 7 and 11 and trained them to explain whether each intervention was useful based on the difference-making principle. The brief single-trial training led them to generate more EIG-based explanations and rely more on EIG in later interventions. If explaining is already a powerful tool in simple cases, sharpening it gives it more might in hard ones.

5.2.2 New way to facilitate intervention

At 4 years of age, children can already learn causal structures by observing someone else's interventions[113,119]. However, even much older children have limited capacities to choose and learn from self-generated interventions. In a minimal causal system with only two nodes (one or both broken), Lapidow and Walker[69] demonstrated that 4- to 6-year-olds selected informative interventions (i.e., testing the node that may *not* be broken) and identified the correct hypothesis. However, this minimal system grossly simplified real-world causal learning problems. To begin, the connection between the two nodes was a given, only that a broken node cannot influence its neighbor through this connection. As a result, structural learning wasn't required. Moreover, since two nodes only form one edge, this problem doesn't shed light on whether child learners laboriously test one edge a time or consider the global structure of the causal system as a whole when testing hypotheses.

When the causal system is slightly more complex (three or more nodes), children often fail to choose informative interventions. McCormack et al.[83] found that 5- to 7-year-olds didn't consistently outperform models that selected interventions at random. Granted, 7- to 8-year-olds performed better than chance, but the chance model didn't capture how much each child used EIG and PTS. In Nussenbaum et al.'s study that employed the EIG and PTS hybrid model[90], it wasn't until 12 years of age that children started to use EIG as their primary strategy. My findings are consistent with the previous study: Five- to 11-year-olds heavily relied on PTS without being trained to maximize EIG (5-year-olds in Studies 1–2; 7- to 11-year-olds in the Explanation condition in Study 3).

Given children's difficulty choosing interventions, it's worth noting that after brief EIG training, 9- to 11-year-olds relied on EIG nearly as much as adults did (children's EIG weight θ was .77 in the Training condition in Study 3; adults' was .77 in the Baseline condition in Study 1). Another key lesson from this dissertation is this: To facilitate intervention selection, we can teach children the abstract principle of good interventions, *i.e.*, they lead to different outcomes under different hypotheses. This process may only take 2 minutes but can profoundly shape children's intervention strategy (9- to 11-year-olds' θ went from .45 in the Explanation condition to .72 in the Training condition).

5.3 Limitations

Several limitations in this dissertation can be immediately addressed in future studies. To begin, after intervening on uninformative light bulbs, some children said *"I don't know which picture is correct. Maybe both?"*. Those who didn't say so explicitly might have also felt uncertain. In future studies, we can ask learners to rate their confidence of the chosen structure based on the intervention outcome. A growing body of studies[32,68,116,121] showed that children's confidence or uncertainty judgments were tightly related to their exploratory behavior. In the case of causal intervention, those

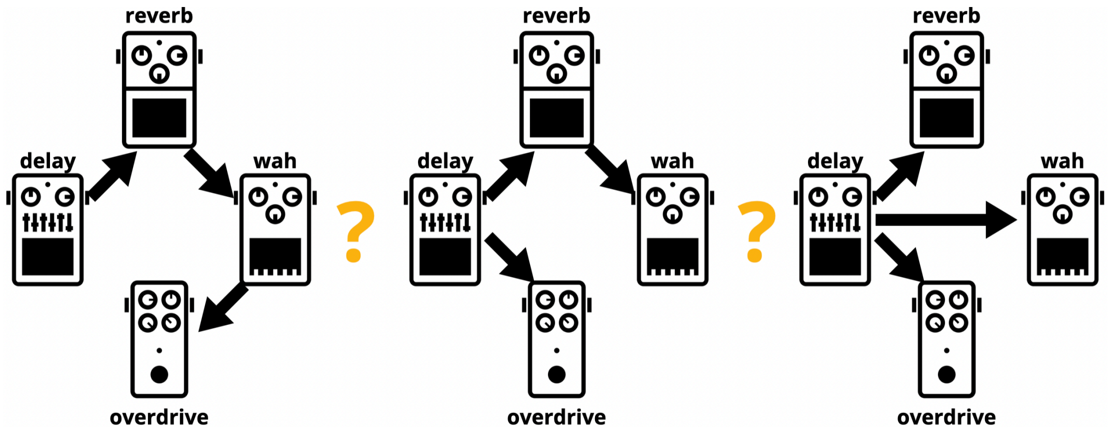


Figure 5.1: A pedal board with delay, reverb, overdrive, and wah pedals connected in an unknown way.

who judge their confidence appropriately may improve more with practice than those who have low confidence after informative interventions or high confidence after uninformative ones.

Another important limitation in this work is the lack of a baseline in Study 3. EIG training increased children's EIG reliance, but it's possible that being prompted to explain can already help 7- to 11-year-olds choose better interventions. Granted, adults didn't intervene differently as a result of self-explaining, but their dominant strategy was already EIG in the Baseline condition, leaving little room for improvement. To examine this question more rigorously, we can test 7- to 11-year-olds in the Baseline condition where they choose interventions spontaneously as well as in the Report condition where they report their choice before intervening on it. If those in the Explanation condition rely more on EIG compared to the Baseline and the Report conditions, it suggests that self-explaining alone can already help older children choose better interventions to some degree.

5.4 Future Directions

This work opens up exiting new directions in learning by experimentation and learning by thinking.

5.4.1 Why did learners use PTS when it's a suboptimal strategy?

The foremost is, why did learners sometimes use PTS that's less optimal than EIG? Why would one turn on a light bulb that activates everything in both pictures? There are many potential reasons.

First of all, PTS users may be "resource-rational"[55,73] learners who strike a balance between the quality of intervention (e.g., measured by EIG) and the computational cost (e.g., measured by the number of "expensive" computations that take all structures into account). PTS is less computationally expensive than EIG, especially as the number of nodes grow. In the four-node causal system introduced in Chapter 1 (Figure 5.1), there are $3^{\binom{4}{2}} = 729$ different ways the nodes could be connected. To identify the correct structure among the 729 possibilities, an EIG user has to consider what each of the four pedals does in each of the 729 structures. A PTS user only has to carry out this expensive computation once, *i.e.*, when finding the node that affects the maximum proportion of links across 729 structures. PTS only takes one more intervention than EIG: The delay pedal that has the highest PTS score doesn't reduce any uncertainty, but the reverb pedal that has the second highest PTS score is informative. In future work, we can manipulate each strategy's computational cost relative to the intervention quality to see whether learners can adapt their strategies.

Another possible reason for using PTS is that learners mistakenly think of uninformative interventions as informative. For instance, if a learner thinks the delay pedal only turns on all four in the rightmost structure, they may believe turning it on might rule out the first two. Coenen and Gureckis[27] found indirect evidence supporting this idea. In their study, if learners chose an uninformative intervention, they shouldn't be able to tell which of two hypotheses equally was correct, yet some showed a strong bias for one over the other. The stronger the bias, the less learners relied on EIG. To test it directly, we can present learners with uninformative interventions and ask them to identify the correct structures. An unbiased learner should be uncertain as to which structure is correct whereas a biased one may endorse one structure despite inconclusive evidence. We can then test them in an intervention selection task (akin to the Light Bulb Game) and examine whether a learner's bias correlates with their mean EIG score and weight of EIG θ later on.

Recently, Lapidow and Walker[70] argued that discriminating between structures isn't the only goal learners have — if their goal is to test causal *stability*, then PTS is the appropriate strategy. For instance, after getting a new key, we often try it on the door a few times, just to make sure it works reliably. On the surface, this behavior seems redundant since we don't gain new information about how the key and the door are causally linked (i.e., it's the key that opens the door, not the other way around), but we still wish to ensure this important connection is stable, so we don't lock ourselves out. In the Light Bulb Game, this reason doesn't apply because causal relationships are deterministic: If light bulb A has an arrow pointing to B, then A can definitely turn on B. In future studies, we can use probabilistic causal relationships and examine whether learners lean more towards PTS.

5.4.2 What does PTS truly entail?

Taking a step back, what goes on in a learner's mind when they use PTS? Formal definitions and learners' explanations can both vary. In most puzzles, different PTS definitions point to the same intervention. In Figure 5.2, different definitions may favor different intervention choices.

In all studies[27,29,85,86,90] that explicitly modeled PTS, a given intervention's PTS score is defined as the maximum proportion (maxProp) of links it can affect across all structures. This choice is somewhat arbitrary: An intervention's PTS score can also be defined as the the maximum number (maxNum) of links it can affect, the mean proportion of links (meanProp), or the mean number of links (meanNum). As shown in Table 5.1, definitions based on the maximum favor the yellow light bulb that could potentially turn on all light bulbs. By contrast, definitions based on the mean assign the highest PTS score to the green light bulb, which always turn on another light bulb. Interestingly, PTS explanations children generated seemed to fall under the two types of PTS. Some said they wanted to turn on the yellow light bulb because *"it may turn on all three of them"* while some chose

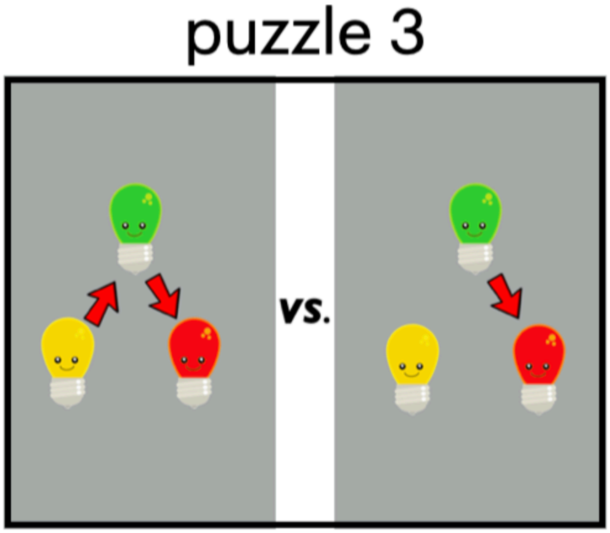


Figure 5.2: An example puzzle where different definitions of PTS may favor different interventions.

light bulb	EIG	PTS (maxProp)	PTS (maxNum)	PTS (meanProp)	PTS (meanNum)
yellow	1	1	2	0.5	1
green	0	1	1	0.75	1
red	0	0	0	0	0

Table 5.1: In the example puzzle, four definitions of PTS score interventions slightly differently.

the green one because *"it always turns on the red"*. In future studies, we can create more puzzles where PTS scores differ by definition and use a Bayesian latent mixture model to capture what type of PTS a learner may be using as well as how much they rely on EIG *vs.* the particular type of PTS.

5.4.3 How else can we facilitate intervention selection?

While effective, training children on EIG maximization is likely not the only way to help them choose better interventions. I propose three ideas below for future researchers to pursue.

First of all, causal intervention involves extensive counterfactual reasoning (*"What if I turn on light A?" "What happens if structure X is true?"*) — engaging learners in counterfactual reasoning may facilitate their intervention selection. In a recent study, Nyhout et al.[91] asked to children to figure out how far a ball would roll down from a ramp with adjustable height and angle. An optimal learner should use the control-of-variables (CV) strategy here, *i.e.*, fixing one variable (e.g., height) while varying the other (angle) to see how it impacts a ball's rolling distance. Seven- to 10-year-olds who received counterfactual scaffolding (i.e., watching a ball rolling down from a ramp and imaging what would happen if the ramp was set higher/lower) were more likely to use the CV strategy than those in the control condition (watching it again). To introduce counterfactual training to intervention selection, we can modify the Training condition in Study 3: Rather than letting children turn on the chosen light bulb in the training trial, we can ask them to imagine the outcome under each hypothesis, *"Shoot, these light bulbs ran out of batteries today! It's alright — let's just imagine that you turned on the [color] light bulb. What would happen if the left picture is the answer? What would happen if the right one is the answer?"*. After the child answers correctly, we can tell them an imaginary outcome and ask them to choose the correct structure accordingly (*"Okay, let's say [outcome] happened. What picture is the answer?"*). It's possible that children benefit more from counterfactual training than if they actually carry out the intervention and observe its outcome.

Another potential way to facilitate intervention selection is asking children to evaluate another person's intervention choice before they choose and learn from their own interventions. Generating informative interventions is computationally expensive whereas judging interventions already chosen is much easier but still invokes the difference-making principle. Seeing someone else activate a light bulb that leads to the same outcome no matter what, do children think it can help identify the correct structure? Forced to choose between an informative and an uninformative intervention, will children go for the former?... After learners have evaluated interventions from a third-person perspective, we can test them in the Light Bulb Game and examine whether engaging in intervention evaluation shapes intervention selection and whether evaluation and selection are correlated.

Last but not least, we can capitalize on learners' mistakes to facilitate learning. After turning on wrong light bulbs, some children expressed regret (*"Oh no, I should've turned on another one!"*) or confusion (*"Huh, how can I know which one is correct?"*). In the infant literature, when an observed outcome violates the baby's expectation, they look longer[2,6], explore accordingly[121], and seek explanations for the unexpected phenomenon[98]. In adult studies, prediction errors between actual and expected outcomes also predict learning[54,82]. In the future, we can seize the moment and ask causal learners to explain an intervention they realize to be incorrect, *"Why don't you think this light bulb can help you find the answer?"*. After that, we can give them an opportunity to correct the mistake, *"If I allow you to turn on another light bulb, which one would you choose? Why?"* We can check to see whether learners who are asked to reflect on an uninformative intervention intervene better in later trials, compared to those who are aware of their mistake but never questioned about it.

5.5 Concluding Remarks

> - Judge: *"Am I never to hear to hear the truth?"*
> - Barrister: *"No, my lord, merely the evidence."*
> — Peter Murphy (1980)

To summarize findings in this dissertation, even in three-node causal systems with only two possible structures, choosing interventions that maximize the expected information gain (EIG) still poses genuine challenges to a wide age range of children, from 5 to 11 years olds. Most children use the suboptimal positive test strategy (PTS) instead of EIG (all conditions in Studies 1-3 except for the Training condition in Study 3) to select interventions. Simply asking children to explain why they choose certain interventions is insufficient to change the strategies they use, even though it improves 5- to 7-year-olds' accuracy at inferring the correct causal structures from the intervention outcomes (Study 2). As a more effective way to facilitate intervention selection, briefly training children on the difference-making principle behind informative interventions (that they lead to distinct outcomes under different structures) helps them select better interventions later on their own. The training effect is stronger in 9- to 11-year-olds than it was in 7- to 8-year-olds. The bottom line is that thinking alone doesn't lead to better experimentation, but training the prepared mind to think might.

References

[1] Angrist, J. D., & Pischke, J.-S. (2008). *Mostly harmless econometrics*. Princeton University Press.

[2] Aslin, R. N. (2007). What's in a look? *Developmental Science*, *10*(1), 48–53.

[3] Austerweil, J. L., & Griffiths, T. L. (2011). Seeking confirmation is rational for deterministic hypotheses. *Cognitive Science*, *35*(3), 499–526.

[4] Baillargeon, R., Kotovsky, L., & Needham, A. (1995). *The acquisition of physical knowledge in infancy*. Oxford University Press.

[5] Baillargeon, R., Scott, R. M., & Bian, L. (2016). Psychological reasoning in infancy. *Annual Review of Psychology*, *67*, 159–186.

[6] Baillargeon, R., Spelke, E. S., & Wasserman, S. (1985). Object permanence in five-month-old infants. *Cognition*, *20*(3), 191–208.

[7] Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., & Mordatch, I. (2019). *Emergent tool use from multi-agent autocurricula.* arXiv. Retrieved from https://

arxiv.org/abs/1909.07528 doi: 10.48550/ARXIV.1909.07528

[8] Bloom, P. (2000). *How children learn the meanings of words.* MIT Press.

[9] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... others (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

[10] Bonawitz, E., Ullman, T. D., Bridgers, S., Gopnik, A., & Tenenbaum, J. B. (2019). Sticking to the evidence? A behavioral and computational case study of micro-theory change in the domain of magnetism. *Cognitive Science*, *43*(8), e12765.

[11] Bonawitz, E. B., & Lombrozo, T. (2012). Occam's rattle: Children's use of simplicity and probability to constrain inference. *Developmental Psychology*, *48*(4), 1156–1164.

[12] 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.

[13] Bramley, N. R., Gerstenberg, T., Mayrhofer, R., & Lagnado, D. A. (2018). Time in causal structure learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(12), 1880–-1910.

[14] Bramley, N. R., Jones, A., Gureckis, T. M., & Ruggeri, A. (2020). Children's failure to control variables may reflect adaptive decision making. *PsyArXiv*.

[15] Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (2015). Conservative forgetful scholars: How people learn causal structure through sequences of interventions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 708–731.

[16] Brockbank, E., & Walker, C. M. (2022). Explanation impacts hypothesis generation, but not evaluation, during learning. *Cognition*, *225*, 105100.

[17] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., … others (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

[18] Carey, S. (1978). The child as word learner. In M. Halle, J. Bresnan, & G. A. Miller (Eds.), *Linguistic theory and psychological reality* (pp. 264–293). MIT Press.

[19] Carey, S. (2009). *The origin of concepts*. Oxford University Press.

[20] Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development*, *70*(5), 1098–1120.

[21] Chi, M. T. (2000). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. *Advances in Instructional Psychology*, *5*, 161–238.

[22] Chi, M. T., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, *13*(2), 145–182.

[23] Chi, M. T., De Leeuw, N., Chiu, M.-H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*(3), 439–477.

[24] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., … Fiedel, N. (2022). *Palm: Scaling language modeling with pathways.* arXiv. Retrieved from `https://arxiv.org/abs/2204.02311` doi: 10.48550/ARXIV.2204.02311

[25] Chu, J., & Schulz, L. E. (2020). Play, curiosity, and cognition. *Annual Review of Developmental Psychology*, *2*, 317–343.

[26] Clement, J. J. (2009). The role of imagistic simulation in scientific thought experiments. *Topics in Cognitive Science*, *1*(4), 686–710.

[27] Coenen, A., & Gureckis, T. M. (2015). Are biases when making causal interventions related to biases in belief updating? In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 411–416). Austin, TX: Cognitive Science Society.

[28] Coenen, A., Nelson, J. D., & Gureckis, T. M. (2019). Asking the right questions about the psychology of human inquiry: Nine open challenges. *Psychonomic Bulletin & Review*, *26*(5), 1548–1587.

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

[30] Coenen, A., Ruggeri, A., Bramley, N. R., & Gureckis, T. M. (2019). Testing one or multiple: How beliefs about sparsity affect causal experimentation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(11), 1923.

[31] Cohen, L. B., & Oakes, L. M. (1993). How infants perceive a simple causal event. *Developmental Psychology*, *29*(3), 421.

[32] Cook, C., Goodman, N. D., & Schulz, L. E. (2011). Where science starts: Spontaneous experiments in preschoolers' exploratory play. *Cognition*, *120*(3), 341–349.

[33] Dehaene, S. (2011). *The number sense: How the mind creates mathematics*. Oxford University Press.

[34] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *2009 ieee conference on computer vision and pattern recognition* (pp. 248–255).

[35] Denison, S., & Xu, F. (2014). The origins of probabilistic inference in human infants. *Cognition*, *130*(3), 335–347.

[36] Denison, S., & Xu, F. (2019). Infant statisticians: The origins of reasoning under uncertainty. *Perspectives on Psychological Science*, *14*(4), 499–509.

[37] Dewar, K. M., & Xu, F. (2010). Induction, overhypothesis, and the origin of abstract knowledge: Evidence from 9-month-old infants. *Psychological Science*, *21*(12), 1871–1877.

[38] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., … Houlsby, N. (2020). *An image is worth 16x16 words: Transformers for image recognition at scale.* arXiv. Retrieved from https://arxiv.org/abs/2010.11929 doi: 10.48550/ARXIV.2010.11929

[39] Edwards, B. J., Williams, J. J., Gentner, D., & Lombrozo, T. (2019). Explanation recruits comparison in a category-learning task. *Cognition*, *185*, 21–38.

[40] Fedorov, V. V. (1972). *Theory of optimal experiments.* Academic Press.

[41] Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*(7), 307–314.

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

[43] Fonseca, B. A., & Chi, M. T. (2011). Instruction based on self-explanation. *Handbook of research on learning and instruction*, *2*, 296–321.

[44] Gershman, S. (2021). *What makes us smart: The computational logic of human cognition.* Princeton University Press.

[45] Gershman, S. J. (2019). How to never be wrong. *Psychonomic Bulletin & Review*, *26*(1), 13–28.

[46] Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278.

[47] Gershman, S. J., Pouncy, H. T., & Gweon, H. (2017). Learning the structure of social influence. *Cognitive Science*, *41*, 545–575.

[48] Goodman, N. D., Ullman, T. D., & Tenenbaum, J. B. (2011). Learning a theory of causality. *Psychological Review*, *118*(1), 110–119.

[49] Gopnik, A. (2012). Scientific thinking in young children: Theoretical advances, empirical research, and policy implications. *Science*, *337*(6102), 1623–1627.

[50] 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.

[51] Gopnik, A., Sobel, D. M., Schulz, L. E., & Glymour, C. (2001). Causal learning mechanisms in very young children: two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology*, *37*(5), 620–629.

[52] Gopnik, A., & Tenenbaum, J. B. (2007). Bayesian networks, Bayesian learning, and cognitive development. *Developmental Science*, *10*(3), 281–287.

[53] Gopnik, A., & Wellman, H. M. (2012). Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological Bulletin*, *138*(6), 1085–1108.

[54] Gottlieb, J., & Oudeyer, P.-Y. (2018). Towards a neuroscience of active sampling and curiosity. *Nature Reviews Neuroscience*, *19*(12), 758–770.

[55] 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.

[56] Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, *116*(4), 661–176.

[57] Gweon, H., Pelton, H., Konopka, J. A., & Schulz, L. E. (2014). Sins of omission: Children selectively explore when teachers are under-informative. *Cognition*, *132*(3), 335–341.

[58] Harrison, E. (1964). Olbers' paradox. *Nature*, *204*(4955), 271–272.

[59] Hegarty, M. (2004). Mechanical reasoning by mental simulation. *Trends in Cognitive Sciences*, *8*(6), 280–285.

[60] Hume, D. (1740/2000). *An enquiry concerning human understanding*. Oxford University Press.

[61] Ichikawa, J. J., & Steup, M. (2018). The Analysis of Knowledge. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2018 ed.). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2018/entries/knowledge-analysis/.

[62] James, W., Burkhardt, F., Bowers, F., & Skrupskelis, I. K. (1890). *The principles of psychology* (Vol. 1) (No. 2). Macmillan London.

[63] Klayman, J., & Ha, Y.-W. (1989). Hypothesis testing in rule discovery: Strategy, structure, and content. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(4), 596–604.

[64] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImagNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*.

[65] Kuhn, D., & Brannock, J. (1977). Development of the isolation of variables scheme in experimental and "natural experiment" contexts. *Developmental Psychology*, *13*(1), 9–14.

[66] Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, e253.

[67] Lampinen, A. K., Roy, N. A., Dasgupta, I., Chan, S. C., Tam, A. C., McClelland, J. L., … others (2021). Tell me why!—Explanations support learning of relational and causal structure. *arXiv preprint arXiv:2112.03753*.

[68] Lapidow, E., Killeen, I., & Walker, C. M. (2022). Learning to recognize uncertainty vs. recognizing uncertainty to learn: Confidence judgments and exploration decisions in preschoolers. *Developmental Science*, *25*(2), e13178.

[69] Lapidow, E., & Walker, C. M. (2020). Informative experimentation in intuitive science: Children select and learn from their own causal interventions. *Cognition*, *201*, 104315.

[70] Lapidow, E., & Walker, C. M. (2021). Rethinking the "gap": Self-directed learning in cognitive development and scientific reasoning. *Wiley Interdisciplinary Reviews: Cognitive Science*, e1580.

[71] Legare, C. H. (2012). Exploring explanation: Explaining inconsistent evidence informs exploratory, hypothesis-testing behavior in young children. *Child Development*, *83*(1), 173–185.

[72] Leslie, A. M., & Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition*, *25*(3), 265–288.

[73] 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.

[74] Lindley, D. V. (1956). On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, *27*(4), 986–1005.

[75] Liquin, E. G., & Lombrozo, T. (2017). Explain, explore, exploit: Effects of explanation on information search. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 2598–2603). Austin, TX: Cognitive Science Society.

[76] Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017). Ten-month-old infants infer the value of goals from the costs of actions. *Science*, *358*(6366), 1038–1041.

[77] Lombrozo, T. (2016). Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences*, *20*(10), 748–759.

[78] Lombrozo, T. (2019). &ldquo;learning by thinking&rdquo; in science and in everyday life. In A. Levy & P. Godfrey-Smith (Eds.), *The scientific imagination* (p. 230-249). Oxford University Press.

[79] Markman, E. M. (1989). *Categorization and naming in children: Problems of induction*. MIT Press.

[80] Markman, E. M. (1990). Constraints children place on word meanings. *Cognitive Science*, *14*(1), 57–77.

[81] Marr, D. (1982). The philosophy and the approach. In *Vision* (pp. 8–29). San Francisco, CA: Freeman.

[82] Marvin, C. B., & Shohamy, D. (2016). Curiosity and reward: Valence predicts choice and information prediction errors enhance learning. *Journal of Experimental Psychology: General*, *145*(3), 266–272.

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

[84] McCormack, T., Frosch, C., Patrick, F., & Lagnado, D. (2015). Temporal and statistical information in causal structure learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(2), 395–416.

[85] Meng, Y., Bramley, N. R., & Xu, F. (2018). Children's causal interventions combine discrimination and confirmation. In C. Kalish, M. Rau, J. Zhu, & T. Rogers (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 762–767). Austin, TX: Cognitive Science Society.

[86] Meng, Y., & Xu, F. (2019). Leveraging thinking to facilitate causal learning from intervention. In A. Goel, C. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (pp. 2338–2344). Austin, TX: Cognitive Science Society.

[87] Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT Press.

[88] Navarro, D. J., & Perfors, A. F. (2011). Hypothesis generation, sparse categories, and the positive test strategy. *Psychological Review*, *118*(1), 120–134.

[89] Nelson, J. D. (2005). Finding useful questions: On Bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, *112*(4), 979–999.

[90] Nussenbaum, K., Cohen, A. O., Davis, Z. J., Halpern, D. J., Gureckis, T. M., & Hartley, C. A. (2020). Causal information-seeking strategies change across childhood and adolescence. *Cognitive Science*, *44*(9), e12888.

[91] Nyhout, A., Iannuzziello, A., Walker, C., & Ganea, P. (2019). Thinking counterfactually supports children's ability to conduct a controlled test of a hypothesis. In A. Goel, C. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (pp. 2488–2494). Austin, TX: Cognitive Science Society.

[92] Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608–631.

[93] Orman Quine, W. v. (1976). Two dogmas of empiricism. In *Can theories be refuted?* (pp. 41–64). Springer.

[94] Pacer, M., & Lombrozo, T. (2017). Ockham's razor cuts to the root: Simplicity in causal explanation. *Journal of Experimental Psychology: General*, *146*(12), 1761–1780.

[95] Pearl, J. (2000). *Causality*. New York: Oxford University Press.

[96] Pearl, J. (2018). Theoretical impediments to machine learning with seven sparks from the causal revolution. *arXiv preprint arXiv:1801.04016*.

[97] Pearl, J., & MacKenzie, D. (2018). *The book of why: The new science of cause and effect.* New York: Basic Books.

[98] Perez, J., & Feigenson, L. (2022). Violations of expectation trigger infants to search for explanations. *Cognition*, *218*, 104942.

[99] Piaget, J. (1955). *The child's construction of reality*. Routledge.

[100] Popper, K. (2005). *The logic of scientific discovery*. Routledge.

[101] Quine, W. V. (1970). On the reasons for indeterminacy of translation. *The Journal of Philosophy*, *67*(6), 178–183.

[102] Quine, W. V. O. (1960). *Translation and meaning*. Cambridge, MA: MIT Press.

[103] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). *Hierarchical text-conditional image generation with clip latents.* arXiv. Retrieved from https://arxiv.org/abs/2204.06125 doi: 10.48550/ARXIV.2204.06125

[104] Reed, S., Zolna, K., Parisotto, E., Colmenarejo, S. G., Novikov, A., Barth-Maron, G., … de Freitas, N. (2022). *A generalist agent.* arXiv. Retrieved from https://arxiv.org/abs/2205.06175 doi: 10.48550/ARXIV.2205.06175

[105] Rhodes, M., Leslie, S.-J., & Tworek, C. M. (2012). Cultural transmission of social essentialism. *Proceedings of the National Academy of Sciences*, *109*(34), 13526–13531.

[106] Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, *26*(5), 521–562.

[107] Ruggeri, A., Swaboda, N., Sim, Z. L., & Gopnik, A. (2019). Shake it baby, but only when needed: Preschoolers adapt their exploratory strategies to the information structure of the task. *Cognition*, *193*, 104013.

[108] Ruggeri, A., Xu, F., & Lombrozo, T. (2019). Effects of explanation on children's question asking. *Cognition*, *191*, 103966.

[109] Rule, J. S., Tenenbaum, J. B., & Piantadosi, S. T. (2020). The child as hacker. *Trends in Cognitive Sciences*, *24*(11), 900–915.

[110] Schulte, O. (2022). Formal learning theory. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy*.

[111] Schulz, L. (2012). The origins of inquiry: Inductive inference and exploration in early childhood. *Trends in Cognitive Sciences*, *16*(7), 382–389.

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

[113] Schulz, L. E., Gopnik, A., & Glymour, C. (2007). Preschool children learn about causal structure from conditional interventions. *Developmental Science*, *10*(3), 322–332.

[114] Searle, J. R. (1982). The chinese room revisited. *Behavioral and Brain Sciences*, *5*(2), 345–348.

[115] Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 379–423.

[116] Siegel, M. H., Magid, R. W., Pelz, M., Tenenbaum, J. B., & Schulz, L. E. (2021). Children's exploratory play tracks the discriminability of hypotheses. *Nature Communications*, *12*(1), 1–9.

[117] Sim, Z. L., & Xu, F. (2017). Learning higher-order generalizations through free play: Evidence from 2- and 3-year-old children. *Developmental Psychology*, *53*(4), 642–651.

[118] Smucker, B., Krzywinski, M., & Altman, N. (2018). Optimal experimental design. *Nature Methods*, *15*(8), 559–560.

[119] Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and bayesian reasoning in preschoolers. *Cognitive Science*, *28*(3), 303–333.

[120] Spirtes, P., Glymour, C. N., Scheines, R., & Heckerman, D. (1993). *Causation, prediction, and search*. MIT Press.

[121] Stahl, A. E., & Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. *Science*, *348*(6230), 91–94.

[122] Steyvers, M., Tenenbaum, J. B., Wagenmakers, E. J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, *27*(3), 453–489.

[123] Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction* (Vol. 1). Cambridge: MIT Press.

[124] Tauber, S., Navarro, D. J., Perfors, A., & Steyvers, M. (2017). Bayesian models of cognition revisited: Setting optimality aside and letting data drive psychological theory. *Psychological Review*, *124*(4), 410–441.

[125] 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.

[126] Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017). Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences*, *21*(9), 649–665.

[127] Vélez, N., Bridgers, S., & Gweon, H. (2019). The rare preference effect: Statistical information influences social affiliation judgments. *Cognition*, *192*, 103994.

[128] Vihman, M. M. (1996). *Phonological development: The origins of language in the child.* Blackwell Publishing.

[129] Walker, C. M., Bonawitz, E., & Lombrozo, T. (2017). Effects of explaining on children's preference for simpler hypotheses. *Psychonomic Bulletin & Review*, *24*(5), 1538–1547.

[130] Walker, C. M., Bridgers, S., & Gopnik, A. (2016). The early emergence and puzzling decline of relational reasoning: Effects of knowledge and search on inferring abstract concepts. *Cognition*, *156*, 30–40.

[131] Walker, C. M., & Lombrozo, T. (2017). Explaining the moral of the story. *Cognition*, *167*, 266–281.

[132] Walker, C. M., Lombrozo, T., Legare, C. H., & Gopnik, A. (2014). Explaining prompts children to privilege inductively rich properties. *Cognition*, *133*(2), 343–357.

[133] Walker, C. M., Lombrozo, T., Williams, J. J., Rafferty, A. N., & Gopnik, A. (2017). Explaining constrains causal learning in childhood. *Child Development*, *88*(1), 229–246.

[134] Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12*(3), 129–140.

[135] Wilkenfeld, D. A., & Lombrozo, T. (2015). Inference to the best explanation (IBE) versus explaining for the best inference (EBI). *Science & Education*, *24*(9), 1059–1077.

[136] Williams, J. J., & Lombrozo, T. (2010). The role of explanation in discovery and generalization: Evidence from category learning. *Cognitive Science*, *34*(5), 776–806.

[137] Williams, J. J., Lombrozo, T., & Rehder, B. (2013). The hazards of explanation: Overgeneralization in the face of exceptions. *Journal of Experimental Psychology: General*, *142*(4), 1006–1014.

[138] Xu, F. (2019). Towards a rational constructivist theory of cognitive development. *Psychological Review*, *126*(6), 841–864.

[139] Xu, F., Spelke, E. S., & Goddard, S. (2005). Number sense in human infants. *Developmental Science*, *8*(1), 88–101.

[140] Xu, F., & Tenenbaum, J. B. (2007a). Sensitivity to sampling in bayesian word learning. *Developmental Science*, *10*(3), 288–297.

[141] Xu, F., & Tenenbaum, J. B. (2007b). Word learning as bayesian inference. *Psychological Review*, *114*(2), 245–272.