Analytic Causal Knowledge for Constructing Useable Empirical Causal Knowledge: Two Experiments on Pre-schoolers

Patricia W. Cheng [1,*] | **Catherine M. Sandhofer** [1] | **Mimi Liljeholm** [2]

[1] Department of Psychology, University of California, Los Angeles

[2] Department of Cognitive Sciences, University of California, Irvine

Correspondence should be sent to Patricia W. Cheng, Department of Psychology, University of California, Los Angeles, CA 90095, USA. E-mail: cheng@lifesci.ucla.edu

Abstract

The present paper examines a type of abstract domain-general knowledge required for the process of constructing useable domain-specific causal knowledge, the evident goal of causal learning. It tests the hypothesis that analytic knowledge of *causal-invariance decomposition functions* is essential for this process. Such knowledge specifies the decomposition of an observed outcome into contributions from constituent causes under the default assumption that the empirical knowledge acquired is *invariant* across contextual/background causes. The paper reports two psychological experiments (and replication studies) with pre-school-age children on generalization across contexts involving binary cause and effect variables. The critical role of causal invariance for constructing useable causal knowledge predicts that even young children should (tacitly) use the causal-invariance decomposition function for such variables rather than a non-causal-invariance decomposition function common in statistical practice in research involving binary outcomes. The findings support the rational shaping of empirical causal knowledge by the causal-invariance constraint, ruling out alternative explanations in terms of non-causal-invariance decomposition functions, heuristics, and biases. For the same causal structure involving candidate causes and outcomes that are binary variables with a "present" value and an "absent" value, the paper argues against the possibility of multiple rational characterizations of the "sameness of causal influence" that justifies generalization across contexts.

Keywords

Causal learning | Causal invariance | Integration functions | Rationality | Cognitive development

1 | The problem to be solved

How do we humans best represent the world so that we are able to achieve desired outcomes? A basic requirement is that the causal knowledge we acquire be *useable*: Whenever we use our past knowledge to achieve a desired outcome (e.g., avoid a certain food to prevent a skin reaction), we are inevitably assuming that the causal knowledge we acquire in a learning context (e.g., meals at home preceding allergic reactions in the past) generalizes to a subsequent application context (lunch at work the next day, food during foreign travel). By *different contexts* with respect to a cause in question, we mean occasions or settings across which the occurrence of (potentially unknown) alternative causes or enabling conditions of the target outcome changes. Unavoidably, an *application context* may differ from the *learning context*. The present paper addresses the issue of whether pre-school children use a causal-induction process that has a rational basis for constructing causal knowledge aimed at being useable.

By adulthood, humans appear to make causal judgments that suggest they (tacitly) assume *causal invariance*—namely, that causes operate the same way across learning and application contexts—both as a default assumption in probabilistic causal induction (e.g., see Buehner, Cheng, & Clifford, 2003; Novick & Cheng, 2004; for reviews and meta-analyses, see Cheng, 1997 and Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008) and as a criterion for revising causal knowledge during generalization to a new context (Liljeholm & Cheng, 2007; Lu, Rojas, Beckers, & Yuille, 2016). They use *probabilistic causal invariance*[1] (Eqs. 1 and 2 in Section 4) rather than *probabilistic associative invariance* (e.g., Jenkins & Ward, 1965; Rescorla & Wagner, 1972). The two measures differ in what "sameness" they measure—sameness of causal influence versus sameness of association—and hence in the conclusions they draw (Cheng, Liljeholm, Clifford, & Ford, 2007; Liljeholm & Cheng, 2007).

The distinction between causal and associative "sameness" goes beyond the distinction between causation and mere association due to intervention or confounding (Spellman, 1996) and is manifested, for example, in the experimental-design principle of avoiding ceiling effects. Imagine two experiments, neither of which has confounding by extraneous factors, and each shows no difference between its experimental and control condition. The same observed "zero association" between the intervention and the outcome in question in the two experiments can have different causal interpretations.[2] Suppose the outcome occurs at an intermediate level in Experiment A but at a ceiling level in Experiment B, and the purpose in both is to test whether the respective intervention produces the outcome. In that case, whereas the "zero association" in A warrants the conclusion that the intervention tested is not effective, the equivalent association in B would not. Untutored adult reasoners conclude noncausality and withhold judgment in the non-ceiling and ceiling situations, respectively, indicating their ability to differentiate between the two conceptions of "sameness" (e.g., Wu & Cheng, 1999; see Zimmer-Hart & Rescorla, 1974, for an analogous capability in rats). A theory that explains reasoners' intuitive withholding of judgment in ceiling situations also explains their greater willingness to

infer causality for data based on interventions (Cheng, 1997, 2000). The theory makes the causal-invariance assumption.

If the concept of causal invariance is essential to the construction of useable causal knowledge, as we and colleagues have argued (Cheng, Liljeholm, & Sandhofer, 2013; Cheng & Lu, 2017; Park et al., 2022), we would expect young children to use it similarly as adults do. A large body of literature on children's causal reasoning shows that children are able to reason causally from a young age (e.g., Cook et al., 2011; Gopnik et al., 2004; Gweon & Schulz, 2011; Kushnir & Gopnik, 2005, 2007; Schulz, 2012; Schulz, Bonawitz, & Griffiths, 2007; Sobel, Tenenbaum, & Gopnik, 2004). For example, children are able to transfer causal knowledge to a novel instance within the learning context (Schulz et al., 2007). They are also able to learn deterministic causal relations and transfer the acquired relation to similar novel variables within the same context (Lucas, Bridgers, Griffiths, & Gopnik, 2014). Moreover, children's self-directed hypothesis-testing behavior, rather than being interpreted as a confirmation bias, may be viewed as having the goal of assessing the scope of invariance of causal hypotheses (Lapidow & Walker, 2021; also see Legare, 2012; Legare et al., 2010). However, to our knowledge, there has not been previous work that tested whether children use causal-invariance functions (Eqs. 1 and 2) in either of their two roles as a default decomposition function during causal induction or as a criterion for revising causal representation during generalization to a new context.

Our present paper reports two experiments that address the computational-level issue (Marr, 1982) of how it is possible to induce useable causal knowledge by testing the two roles of the concept of causal invariance in pre-school-age children. The stimuli in the experiments involved binary cause and effect variables because such variables most readily test causal invariance against associative invariance (e.g., see Liljeholm & Cheng, 2007; Park et al., 2022). Although children's ability to (implicitly) compute statistical regularities has been well documented (e.g., Saffran, Aslin, & Newport, 1996; Denison & Xu, 2010), that ability does not predict whether children compute causal rather than associative invariance (e.g., Jenkins & Ward, 1965; McCullagh & Nelder, 1989; Rescorla & Wagner, 1972). Our experiments do not seek to demonstrate children's impressive computational abilities; nor do they seek to show generalization behavior in general. Their purpose instead is to provide evidence specifically for the kind of generalization crucial to formulating useable causal knowledge, in support of our thesis that the desire to construct useable causal knowledge shapes the resulting knowledge and does so in a more embedded way than do attention and effort.

2 | Why analytic knowledge of causal-invariance decomposition functions shapes causal knowledge construction

To understand why knowledge of causal-invariance functions shapes causal knowledge construction, consider situations in which (1) background causes may be present, (2) background causes and enabling conditions may vary from context to context, and (3) the set of candidate causes under consideration may not include any that generalizes sufficiently well across

contexts. Under these conditions, the reasoner needs a signal to indicate when to look beyond their current candidate set.

Natural settings often hold these challenges. When we want to infer what cures an illness, for example, the illness must have some non-zero probability of occurring due to some background generative cause. The illness may be more or less prevalent in different contexts (e.g., countries). And it need not occur across all individuals, suggesting that background interacting factors or preventive causes may be present. A further challenge is that our initial parsing of events to isolate distinct candidate causes may not yield predictions that generalize to application contexts; the generalizability of the acquired causal knowledge is a matter of degree (Woodward, 2000, 2010). We may encounter occasions on which a relation that we have previously found to be generalizable unexpectedly fails to hold; for example, on a trip up a tall mountain, we find that eggs boiled for the usual amount of time remain uncooked. The reproducibility crisis in medical and psychological research is a reminder of frequent failures to generalize even for costly, planned investigations (e.g., Challenges in Reproducible Research, 2016; Ioannidis, 2005; Open Science Collaboration, 2015), not to mention for everyday causal inference. The need to go beyond one's current set of candidate causes is ever present.

When background causes are present, inference regarding the power of a candidate cause *c* to influence an outcome requires an assumption regarding how the observed outcome is *decomposed* into the influences by *c* and other (e.g., background) causes. The fact that the "no confounding" condition is a standard principle in experimental design attests to the universal need for decomposition that teases apart the influence of *c* from that of background causes. Our implicit adherence to this condition in everyday causal inference, as indicated by the privileged status we confer on manipulation by free will (e.g., see Cheng, 1997; Gopnik, 2009; Kushnir & Gopnik, 2005; Lagnado & Sloman, 2004; Sloman & Lagnado, 2005; Woodward, 2003), is further testament to this need. The decomposition involves an assumption because the contributing causal relations are inherently not differentiable by observation (Hume, 1739/1987, 1777/1975): every observed outcome is the outcome due to the totality of its causes (see Ichien & Cheng, 2022, for a discussion). The functions characterizing the decomposition are often called *decomposition* or *integration functions* (the two terms refer to mathematically equivalent inverse operations).

Given that the goal of causal learning is to attain useable causal knowledge, and the reasoner starts with a necessarily limited set of candidate causes, there would be a need for a signal to indicate when they should revise their causal knowledge and extend their current candidate set. Selecting the best among a fixed set of candidates will not suffice as a learning strategy. Failure to generalize is a signal that serves that goal. For example, when one is puzzled to find that, on the tall mountain, eggs fail to cook in boiling water, this deviation from expectation may prompt one to revise their understanding of the boiling temperature of water or even of matter more generally.

Failure to generalize is indicated by a notable deviation from the *expected outcome assuming causal invariance*. For situations involving one or more observed causes and background causes

of an effect in question, that expected outcome is specified by the relevant *causal-invariance decomposition function* given information on the states of the observed causes, the estimated strengths of the causes to produce the effect, and the state of the effect due to (potentially unknown) background causes (Eqs. 1 and 2 for binary cause and effect variables). The expected outcome assuming causal invariance is knowledge of the kind we term *analytic*.

The distinction between analytic and empirical knowledge (cf. Hume's, 1739/1987, fork of knowledge: "relations of ideas" and "matters of fact") is seldom made in the cognitive and developmental literatures[2]. The defining feature of analytic knowledge is that it is *justified by what logically follows* from the meaning of the concepts in question—in the case under discussion, what follows from the concept of "sameness of causal influence regardless of context." Analytic knowledge can be either qualitative (e.g., if person *x* is a surgeon, they are a medical doctor) or quantitative (e.g., if the causal mechanism in question involving binary causes and effects remains unchanged regardless of causes in the background, the decomposition functions in Eqs. 1 and 2 logically follow). In contrast, empirical knowledge is *justified by observations or experience*.[3]

Why go beyond empirical knowledge of how things combine? Experience tells us how forces do combine, but not how forces *would* combine *if they do not change, as if other forces were not there*. For example, suppose November in Geographic Region A is typically cloudy and yet has 0 precipitation; the region started cloud seeding and observed 6 snowy days that month. Geographic region B, where November typically has 10 snowy days, also started cloud seeding. Knowledge based on observation/experience in Region B tells us that cloud seeding and natural causes of snowfall in combination resulted in 20 snowy days there that month. What it does not tell us is how many snowy days to expect if the two causes of snowfall have invariant/independent influences. Assuming that the influence of cloud seeding estimated in Region A generalizes across regions, analytic knowledge gives the answer: 14 days (on 2 of those days, snowfall is expected to be due to both causes; Eq. 1). Without this answer based on analytic knowledge, one would be unable to tell that having 20 snowy November days in Region B indicates a synergistic relation between cloud seeding and natural causes of snowfall.

To our knowledge, there has not been previous work assessing whether or not children reason in a way consistent with a distinction between analytic and empirical knowledge. That distinction with regard to the concept of causal invariance enables the assessment of deviation from analytic causal invariance (in the "zoo" context in both of our studies). The observed deviation (in the "zoo" context in Study 1 but not Study 2) should prompt the children to revise their causal knowledge.

2.1 | Why causal invariance as a default assumption in causal induction is essential for formulating useable causal knowledge

The default and revision-criterion roles of causal invariance in the learning and application contexts, respectively, may be regarded as two sides of the same coin: which context happens to be "learning" and which "application" is incidental. It should be clear that successfully predicting

the outcome in a new context means that the influence of cause *c* on effect *e* inferred in the learning context is *independent* of influences from background causes in the new context (i.e., the inferred influence remains *unchanged* across contexts). However, a causal relation that has been found to successfully generalize to an application context can be "tested" in the context of the original inference, flipping the "learning" and "application" context labels (Cheng & Lu, 2017; Ichien & Cheng, 2022). Successful generalization therefore must mean that, conversely, the influence of *c* is independent of the influence of background causes in the original learning context (now the application context). In other words, for any causal mechanism in the world that remains unchanged across two contexts, the desired inferred causal relation that generalizes to an application context could not have resulted from the actual data in the learning context *unless* causal invariance is the decomposition function during learning (recall that decomposition and integration are inverse operations).[4]

The sameness of causal influence implies different mathematical functions depending on the *form* of the cause-and-effect variables (e.g., whether they are binary or continuous) rather than on their specific content (e.g., tobacco smoking and lung cancer). For a given variable type, the causal-invariance and *causal-interaction* (i.e., non-causal-invariance) decomposition functions yield different causal conclusions (e.g., see Buehner et al., 2003; Cheng, 1997; Liljeholm & Cheng, 2007; Lu et al., 2008). In view of the learning and application contexts being transposable, inducing causal relations under a causal-interaction decomposition function (e.g., an associative invariance function) would not serve the goal of acquiring generalizable causal knowledge.

Note that the vastness of the search space of possible causal models changes the nature of the causal-induction problem: An infinite search space renders the use of the concept of causal invariance essential rather than merely helpful. A basic tenet of cognitive science—that our perception and conception of reality are our representations—implies that the search space of the representation of reality is infinite. In an infinite search space, an exhaustive evaluation of the possible causal models is not only practically infeasible (cf. Bramley, Dayan, Griffiths, & Lagnado, 2017) but inherently impossible. In that space, deviation from the outcome predicted/expected assuming causal invariance—along with the constraints of logical consistency and parsimony—serves as a signal for a need to step outside one's current set of candidate causes.

2.2 | A non-causal-invariance decomposition function as a revision criterion

What if the need for revision is signaled instead by deviation from a causal-interaction criterion, a deviation from the case where candidate *c*'s influence on target effect *e* is *expected to vary* depending on the state of the background causes? In that case, there would be a deviation from

expectation—signaling a need to revise causal knowledge—when the influence of *c* in fact successfully generalizes across contexts.

At the same time, no deviation from that expectation would confirm that *c* interacts with background causes (its inferred influence therefore should not generalize across contexts). But, no deviation from expectation means no signal to revise (i.e., the ungeneralizable relation would be kept). Given this inversion of the proper signal to revise, in view of the infinite search space of possible representations of reality, the acquired causal knowledge cannot be expected to hold when applied or to replicate when further tested.

3 | Motivation for study 1

If the essentiality of the concept of causal invariance is correct, we would expect even young children to use the concept. Our two studies on pre-school-age children tested their use of a causal-invariance versus non-causal-invariance criterion for revising their causal knowledge induced from observations presented in the study. Our experimental material in Study 1 was designed so that deviation from the expected outcome (how many animals would have red dots on their face) computed based on causal-invariance versus causal interaction as the decomposition function yield opposite recommendations for an action to achieve a goal (getting rid of red dots).

The materials in both studies involve binary preventive candidate causes (the presence vs. absence of "treats") of a binary outcome (the presence vs. absence of red dots on an animal's face in our story). They concern a scenario involving the evaluation of the effects of two treatments for removing (or preventing) an undesirable outcome, in order to decide which treatment best removes the outcome. Generalizing across contexts in the scenario involves generalizing from a farm context to a zoo context. The farm and zoo have quite different causal contexts in that they differ substantially in the *base rates* of the outcome, the probability of the outcome occurring due to background causes (see Tables 1 and 2). As we explain presently, the event frequencies in Study 1 give rise to different recommendations for action depending on the decomposition function adopted, and the divergence in the recommendations does not diminish with increased sample size.

The children listened to an interactive story titled, "How to get rid of red dots?" The story concerns two brothers—a farmer and a zookeeper—who noticed that some of their animals had red dots on their faces (a fictitious novel binary outcome). The red dots were non-threatening in that they were not a sign of illness. The children were told, "The animals didn't seem sick at all, but the red dots made them look kind of funny." The brothers heard that two "really tasty" and healthy treats, one a *grain* and the other *leaves*, might make the red dots go away. They decided to figure out whether the treats work. First, they visited the farm, where there were 10 farm animals and fed the grain treat to every farm animal; later, they visited the zoo, where there were 10 zoo animals, and fed both the grain and the leaves treat simultaneously to every zoo animal.

Table 1 displays the pattern of event frequencies at the farm and the zoo for Study 1. For both studies, the critical transfer question is as follows: To relieve red dots on new farm and zoo animals, if one has to choose one and only one treat, what is one's best bet on which treat to use grain or leaves? (The transfer question can be equivalently stated in terms of an interaction with something in the background. Both variants address whether one's initial causal belief regarding relieving red dots requires revision.)

Table 1. *Event frequencies for Study 1*

	Farm	**Zoo**
Intervention	grain only	grain & leaves
Pre intervention: animals with dots	9/10	4/10
Post intervention: animals with dots	6/10	1/10
Number Cured	3	3
Fraction Cured	3/9	3/4

Regardless of how "sameness of influence" is defined (i.e., common across associative and causal models), the rationale underlying the choice is as follows: Assuming the grain operates the *same* way across contexts (i.e., farm and zoo), then if the influence of the *intervention*(grain at farm vs. both treats at the zoo) is observed to be the same across contexts, one's best guess would be that leaves had no influence—grain alone would already explain the observations. But, if the influence of the intervention varied across contexts, one would attribute that difference to leaves.

The difference in attribution between the two approaches rests on how "sameness" is defined, causally or associatively. As we explain in Sections 4 and 5.2 for Study 1, deviation based on causal invariance predicts revising the causal network involving three causes (the preventive grain treat and the generative farm and zoo contexts) to include a stronger fourth cause (the leaves treat). The revision leads to the recommendation to give leaves to the new animals. Non-causal-invariance functions—the linear function adopted in associative learning models in psychology (e.g., Jenkins & Ward, 1965; Pearce, 1987, 1994; Rescorla & Wagner, 1972) and the generalized linear model in standard logistic regression (Fienberg, 1980/2007; McCullagh & Nelder, 1989; Wickens, 1989)—predict no deviation from their expected outcome, hence no revision from the three-cause network (i.e., no attribution to leaves). Grain is, therefore, the recommendation according to associative models.

The pattern of outcome frequencies in Table 1 was constructed to address rationality in causal-knowledge construction, not only for psychological models of causal induction but also for their scientific counterparts. The essentiality argument predicts that even young children would adopt a causal-invariance function and recommend a more rational action than associative psychological or "normative" models.

3.1 | Relation to empirical knowledge of integration functions and why assuming causal invariance as a default is not wishful thinking

Several lines of work closely related to ours have shown that people learn or use various integration functions depending on the available data or prior causal knowledge. Waldmann (2007) found that people predominantly integrated influences from two causes using either a linear or an averaging function depending on their knowledge regarding the outcome in question. For example, participants estimated pleasure in the taste of a mixture of two drinks to be the average of the pleasure due to the taste of the individual drinks. Similarly, adults, children, and rats presented with observations that are best described by various (interaction or invariance) integration functions have been shown to be able to learn such functions and, moreover, to generalize the better-fitting one to novel variables with similar content (e.g., Beckers, De Houwer, Pineño, & Miller, 2005; Beckers, Miller, De Houwer, & Urushihara, 2006; De Houwer, Beckers, & Vandorpe, 2005; Lovibond, 2003; Lovibond, Been, Mitchel, Bouton, & Frohardt, 2003; Lucas & Griffiths, 2010; Lucas et al., 2014; Mehta & Williams, 2002; Melchers, Lachnit, & Shanks, 2004; Shanks & Darby, 1998; Urcelay & Miller, 2010; Wheeler, Beckers, & Miller, 2008). This substantial body of work all study the generalization of acquired *empirical*(data-based) integration functions defined on a set of given variables.

In contrast, our work examines the role of causal-invariance functions as analytic knowledge, operating both as a default assumption and a revision criterion in causal-knowledge construction, motivated by the (tacit) aspiration of formulating useable causal knowledge in a reality that is accessible only via representations of it in our mind. To explain the relation between our work and previous work on integration functions, we and our colleagues make two distinctions: (1) a part–whole distinction, between a "*whole*" cause (elemental or complex) and an *interactive component* within a whole cause and (2) a distinction we made earlier between analytic and empirical knowledge (Cheng et al., 2017; Cheng & Lu, 2017; Park et al., 2022). To enable prediction across contexts, the aim of causal-knowledge construction is to formulate whole causes (elemental or complex) that are teased apart from, that do not interact with, other causes (e.g., whole causes in the background). In other words, whole causes are not incidentally independent of other whole causes; we revise causal knowledge—for example, by formulating a conjunctive cause composed of all known interacting components (e.g., Lucas et al., 2014; Lucas & Griffiths, 2010; Melchers et al., 2004; Shanks & Darby, 1998) or by introducing a new whole cause (e.g., in the zoo context in Study 1)—to attain causal invariance between whole causes. Our work and work on empirical integration functions can thus be viewed as complementary if the (unstated) aim of the acquired integration functions is to characterize whole causes that have independent influences on the outcome.

The conception of gravitational forces introduced by Isaac Newton is an exemplar of the distinction between whole causes and their interactive components. For any two celestial bodies *A* and *B*, the empirical integration function describing the interaction among component factors—the masses of bodies *A* and *B* and the distance between them—specifies the gravitational pull between *A* and *B*, *independently* of the masses of all other celestial bodies and the distances from them. That is, gravitational forces do not interact with each other; they are whole causes. The resulting force on a celestial body is the vector sum of the so determined

gravitational forces from all other celestial bodies on the body, and vector sum is the relevant analytic causal-invariance integration function assuming Newtonian time and space.

Placing our two pre-schooler studies in the context of the aspiration of formulating whole causes makes explicit why adopting causal invariance as a default decomposition function is not wishful thinking: deviation from (analytically defined) causal invariance serves as a rational signal for revising causal hypotheses toward restoring invariance in terms of whole causes. The invariance constraint is a crucial part of the human mind's solution to the problem of inducing useable knowledge in the infinite search space of causal representations. The history of science has shown that scientists adopt that constraint: in the face of deviations from causal invariance, scientists seek to revise their representations to restore causal invariance rather than give up the aspiration (Kuhn, 1962/2012; Woodward, 2000).[5] Do pre-school children share that required constraint? To keep our experimental materials simple for the purpose of testing the children, the candidate causes are either whole causes or noncausal.

4 | Causal-invariance decomposition functions for binary variables

The causal-invariance functions for two binary causes of a binary outcome—a candidate cause of the outcome and alternative causes as a group—are as follows (e.g., Cheng, 1997; Pearl, 1988). Below are logically consistent functions for, respectively, generative and preventive candidate causes.

For a candidate cause c that potentially *generates* effect e and does so independently of alternative causes in a given context, denoted a as a group, the probability of observing e is given by a "noisy-OR" decomposition function:

$$P(e = 1|c; w_a, q_c) = q_c \cdot c + w_a - q_c \cdot c \cdot w_a \quad (1)$$

where $c \in \{0,1\}$ denotes the absence and the presence of candidate cause c, $e \in \{0,1\}$ denotes the absence and the presence of effect e, q_c represents the generative power of the candidate cause c, and w_a represents the probability that e occurs due to all background causes, known and unknown.

Likewise, for a candidate cause c that potentially *prevents* effect e independently of a, the probability of observing e is given by a "noisy-AND-NOT" decomposition function:

$$P(e = 1|c; w_a, p_c) = w_a(1 - p_c \cdot c) \quad (2)$$

where p_c is the preventive causal power of c. These "*noisy-logical*" decomposition functions (including noisy-OR, noisy-AND-NOT, and compositions of them, terminology due to Yuille & Lu, 2008), under the assumption that there is *no confounding* [i.e., when $P(a = 1|c = 1) = P(a = 1|c = 0)$], imply respectively equations for estimating q_c and p_c. The equation for estimating p_c (Cheng, 1997; Sheps, 1958) is:

$$p_c = \frac{P(e = 1|c = 0) - P(e = 1|c = 1)}{P(e = 1|c = 0)} \quad (3)$$

COVID vaccine efficacies (e.g., 94.1% for Moderna's mRNA-1273 vaccine), for example, are instantiations of Eq. 3 (NIH News Release, December 30, 2020).

Our experiments test pre-schoolers' use of noisy-logical functions in their role as analytic knowledge of causal invariance for binary cause and effect variables.

5 |Pre-schooler experiments, replications, and adult analogs

Do young children use causal-invariance functions as their default decomposition function and criterion for causal-knowledge revision, as the essentiality of causal-invariance functions predicts? The use of the functions is likely via an implicit process. Our two studies with pre-school children tested our causal-invariance hypothesis against alternative hypotheses, including the linear-decomposition function tested in Liljeholm and Cheng (2007), the generalized-linear decomposition function in logistic regression, and simple heuristic and biases. The cover story and materials in Study 1 are as sketched in Section 3 and detailed in Section 5.3.

One of the non-causal-invariance functions Study 1 tested is the *linear function* adopted in associative learning models in psychology (e.g., Jenkins & Ward, 1965; Pearce, 1987, 1994; Rescorla & Wagner, 1972). The linear decomposition function states that the observed value of the outcome (the probability of an animal having red dots) is explained by the arithmetic sum of the influences of the individual causes present (e.g., the sum of the strengths of the grain treat and background causes at the farm in our story).

Another non-causal-invariance function tested concerns the logistic regression model, probably the most commonly used statistical model for analyses involving multiple causal influences on a binary outcome. Logistic regression is a *generalized linear model* used for predicting the probability of the occurrence of a binary outcome by fitting data to a logistic function of a linear combination of the input variables (Fienberg, 1980/2007; McCullagh & Nelder, 1989; Wickens, 1989). This model amends the linear decomposition function with a logistic scale transformation to render the linearity mathematically coherent. It is widely used for evaluating causal hypotheses in medical and business research, where the hypotheses often involve a binary outcome (e.g., whether or not a bone is fractured, a tumor is malignant, a woman is pregnant, a patient survives, a reader subscribes to a magazine, a shopper buys an item). The argument regarding the essential role of causal-invariance decomposition functions during learning implies that the logistic regression model is not a rational default choice if acquiring generalizable causal knowledge (i.e., if replicability) is the goal. (See Ichien & Cheng, 2022, for an interpretation of logistic regression as a less parsimonious model in which the candidate causes exert independent influences on a mediating ratio-scale continuous variable.)

5.1 | Relation to previous studies on causal learning

Our studies go beyond previous studies on causal learning in adults in two ways: First, because pre-school-age children have fewer acquired normative skills at their disposal (e.g., they are unable to solve the inference problem algebraically), our studies rule out the use of those skills and provide clearer evidence for a natural implicit causal-induction process that aims at formulating useable causal knowledge (Section 2). Second, by avoiding extreme probabilities (0 or 1), the outcome frequencies in Study 1 extend the theoretical scope of our analysis beyond psychological models to include generalized linear models in "normative" associative statistics (e.g., Fienberg, 1980/2007; McCullagh & Nelder, 1989; Wickens, 1989). Causal-invariance and associative-invariance decomposition functions predict not only a reversal in the ordering of the strengths of two candidate causes in the study but also a difference in the acquired causal structure.

5.2 | Predictions for the action to recommend

Whereas the causal-invariance function predicts recommending leaves, models adopting a linear or generalized linear decomposition function—whether frequentist or Bayesian—recommend using grain. A prerequisite for testing these hypotheses is children's willingness to generalize across contexts. Whether this prerequisite is met can be discerned: If children are completely unwilling to generalize across contexts, they would withhold judgment or choose randomly. Here we briefly sketch inferences according to the causal-invariance, linear, and generalized linear models (for prediction details, see the Appendix).

5.2.1 | Predictions according to causal invariance

First, according to the noisy-AND-NOT decomposition rule (predicting responses by Bayesian maximum-likelihood estimates of causal strengths under the preventive causal power assumptions; Cheng, 1997; Griffiths & Tenenbaum, 2005; Tenenbaum & Griffiths, 2001), the outcomes at the farm suggest that the grain removes red dots in a farm animal with a 1/3 probability. To see the rationality underlying this estimate and the recommendation to pick leaves, it may be helpful for the reader to consider the best answer to a simple question for the farm and zoo animals: For the 10 farm animals (see Table 1), with what probability must the treatment (i.e., grain) remove red dots in *each* animal, so that six of the initial nine animals with red dots continue to have red dots? Given that there is no individuating information about the animals, and the treat ingested by an animal cannot "know" what the treat does in other animals (the independent-trials assumption), it should be clear that the best answer is 1/3. Likewise, for the 10 zoo animals, with what probability must the treatment (i.e., both grain and leaves) remove red dots in *each* animal so that only one of the initial four zoo animals with red dots continue to have red dots? The clear answer is 3/4. These intuitive answers follow from Eq. 3.

Now, assuming the causal invariance of grain across the farm and zoo animals, grain would be expected to remove red dots in only 1/3 of the four zoo animals with red dots, considerably less

than 3/4 of the four due to the treatment at the zoo. The leaves then must be explaining the large difference between the expected and the observed outcome. The causal-invariance function, therefore, predicts recommending leaves for the new animals. Note the use here of deviation from the invariance of grain as a criterion for revising causal beliefs by introducing leaves as a whole cause to restore causal invariance.

5.2.2 | Predictions according to the linear and generalized-linear models

In contrast, according to the linear decomposition function (e.g., Jenkins & Ward, 1965; Rescorla & Wagner, 1972; Bayesian maximum-likelihood estimates of causal strengths using the linear decomposition function, Tenenbaum & Griffiths, 2001; Griffiths & Tenenbaum, 2005), the "treatment" at the farm and that at the zoo each subtracts three animals with red dots from the total number of animals with red dots before the treatment at each place. Accordingly, the addition of the leaves treat at the zoo has no observed effect. Therefore, this function concludes that grain treat does all the work and recommends giving it to the new animals.

Logistic regression concurs with the recommendation by the linear psychological models. We explain this prediction visually using Fig. 1. The vertical axis is *P*(red dots), the probability of an animal having red dots; the horizontal axis is the sum of the weights of the predictor variables. For the outcome frequencies in Table 1, because the pattern of events is symmetrical around the probability of .5, according to the logistic regression model the same *reduction* in *P*(red dots)—namely, 3/10—occurs at the farm and the zoo at symmetrical segments of the logistic curve (see vertical dashed lines in Fig. 1, green in an online pdf). Therefore, the grain (see horizontal heavy dashed lines, brown in an online pdf)—which explains the reduction in *P*(red dots) at the farm—explains the entire reduction at the zoo as well.[6] That is, logistic regression detects no influence at all from leaves, either by itself or in an interaction. As the figure makes clear, increasing the sample size does not change this conclusion.

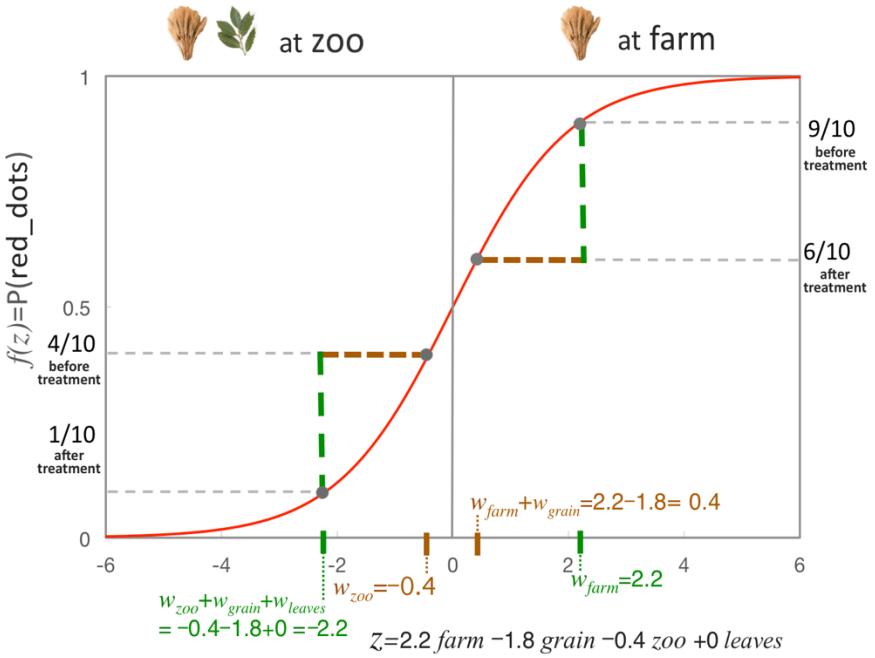

***Fig.* 1** A schematic explanation of the probability of the outcome according to logistic regression: the probability of an animal having red dots at the farm and the zoo, before and after the respective interventions in the scenario, as a logistic function of the sum of the weights—w_{farm}, w_{zoo}, w_{grain}, and w_{leaves}—of the four predictor variables *farm, zoo, grain,* and *leaves* that are present on a trial (e.g., *zoo* = *leaves* = 0 at the farm). The causal strength of grain is represented by a horizontal heavy dashed line (brown in an online pdf); the change in the probability of having red dots due to an intervention is represented by a vertical dashed line (green in an online pdf).

Comparing the predictions by the decomposition functions, note that the linear and logistic functions predict the outcome at the zoo with perfect accuracy while assuming *fewer* causes (grain alone) than does the noisy-AND-NOT function (both grain and leaves). Thus, even perfect prediction accuracy does not imply generalizability, because the prediction rests on assuming a non-causal-invariance function (by our analysis in Section 2). Moreover, choosing leaves cannot be explained by a preference for parsimony.

Next, we report the method and results from four experiments: Study 1 and Study 2, a replication of Studies 1 and 2 with random assignment across studies, and an adult analog of the replication.

5.3| Study 1

5.3.1 | Method

5.3.1.1 | Participants

The participants were 29 children (16 female) with a mean age of 3.42 years (range 2.61–4.84 years, *SD* = 0.60 years). One additional child was excluded for failure to complete the task. Children were recruited from pre-schools in Los Angeles, CA. All children were fluent speakers of English and were learning English as a primary language.

5.3.1.2 | Procedure

As mentioned, children first watched and listened to the picture-book story about the farm and zoo animals with and without red dots on their faces. The farm animals received a grain–treat intervention, and the zoo animals received a simultaneous grain-and-leaves-treat intervention. At the end of the story, children were shown new farm and zoo animals and asked to choose between two potential interventions.

The experiment involved a long sequence of trials (40 animal pictures during the learning sequence), and no trial was repeated. Because inattention to even two trials out of the 40 may lead children to choose a treat randomly, the experimenters were instructed that children's attention was required on every single trial. They were trained to engage and monitor the child's attention for the duration of the experiment.

5.3.1.3 | Storybook task

The task was presented in a child friendly format, as an interactive storybook titled "How to get rid of red dots." The "reader" of the book, and of course the children, were blind to any hypotheses of the study. Children were read the title and the following cover story with accompanying pictures:

"Once upon a time there were two brothers (on the page was a picture of a farmer and a zookeeper; see Fig. 2), one was a farmer and the other a zookeeper. The two brothers loved their animals very much and took very good care of them. One day, the brothers noticed that some of their animals had red dots on their faces (the page showed big red dots)."

After being reassured that the animals were not sick, the children were told about the two treats (the page showed a grain treat and the leaves treat) and were asked to determine their efficacy for making the red dots "go away."

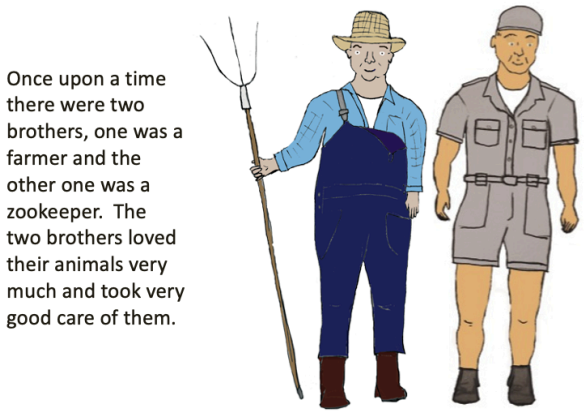

Fig. 2: Illustration of the farmer and the zookeeper in the cover story.

"The two brothers decided to figure out whether the treats work. First, they went together to the farm. Then, they went over to the zoo. Let's look at what happened and see if YOU can figure out if the grain treat makes the red dots go away and if the leaves treat makes the red dots go away."

The farm context and the zoo context were presented separately, and the change in context was highlighted and emphasized. As mentioned, the farm animals received the grain intervention only, whereas the zoo animals received the grain and the leaves intervention in combination. At the end of the trials at the farm, the experimenter said to the child, "Ok, you saw what the grain did to the red dots on the farm animals," then took a long pause to allow the child to make an inference about the grain. She then reminded the child, "At the end of the story, remember, you're going to tell the brothers whether the grain or the leaves works best." At the transition to the zoo, the picture of the two brothers appeared again while the experimenter said, "Now, let's go to the zoo."

Fig. 3 depicts examples of the pre- and post-intervention pictures that children saw. For each animal, because it was critical for children to attend to (1) the presence or absence of red dots and (2) the administered intervention, those aspects of the story were interactive. For example, children were told "Here is a cow before it ate anything today" and then were asked, "Does this cow have any red dots?" Children's responses were acknowledged (e.g., "You're right he does have red dots"). Children were then handed a cutout of the treat to feed to the cow. After the feeding, children were asked to make a prediction (e.g., "Do you think the cow will have red dots on its face now that it ate the grain?") Following the child's reply, the experimenter said, "Let's see!" and showed the picture of the animal with the treat inside its belly, and the presence or absence of red dots was noted regardless of how the child replied (e.g., "Look no more red dots!"). This procedure for a trial was repeated with all 20 animals.

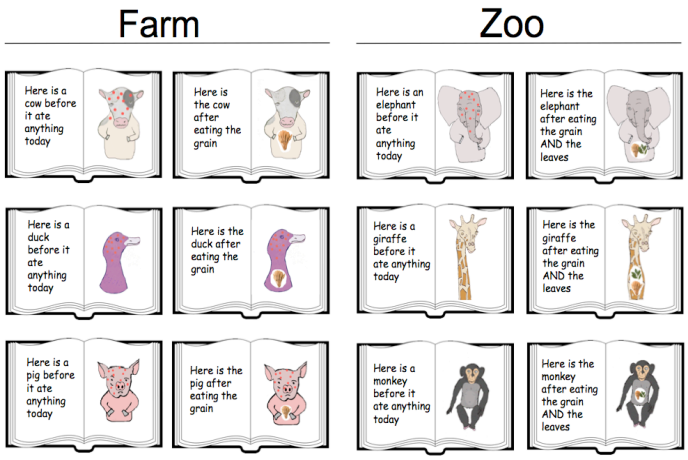

Fig. 3: Examples of the pre- and post- intervention pictures.

5.3.1.4 | Treat selection

The critical test was presented to children at the conclusion of the story. Children were shown two new animals (one farm and one zoo animal) with red dots on their faces and were asked to select only one of the treats, either the grain or the leaves, to make the animals' red dots go away.

5.3.1.5 |Event frequency

The event frequencies for this study are as specified in Table 1. To control for primacy and recency effects, the first trial at the farm and the zoo showed the same event type (an animal showing red dots before treatment but *no* red dots after treatment); likewise, the last trial at the two locations showed the same event type (an animal showing *no* red dots both before and after treatment). (A replication of the study randomized trial order.)

5.3.2 |Results

Children were attentive during the storybook reading and rarely responded incorrectly about the presence or absence of red dots. Across all children and all red-dot observation queries, there were seven initial incorrect responses (out of 580 total queries). For these seven responses, children were corrected (e.g., "Look, here are red dots") and queried again.

The critical result concerned which treat children selected to make the animals' red dots go away. As Fig. 4 shows, children overwhelmingly chose the leaves X^2 (1, N = 29) = 12.4, p = .0004, suggesting that children's responses fit with the noisy-AND-NOT function rather than with the linear function or logistic regression. They did so despite the linear rule's relative arithmetic simplicity and its perfect accuracy in predicting the outcome at the zoo using fewer causes.

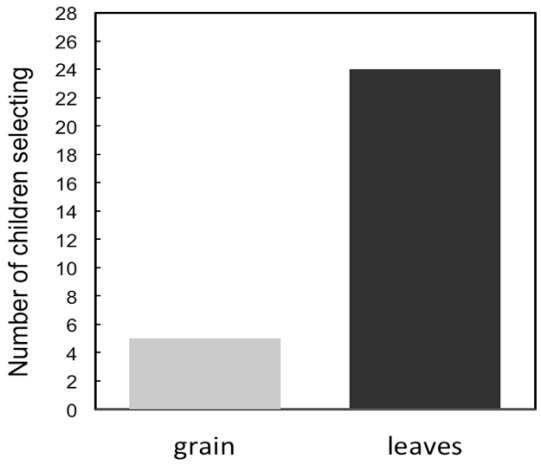

Fig. 4: Results from Study 1 depicting the number of children selecting the grain treat versus the leaves treat.

5.3.3 | Discussion

An enabling condition for our findings is that our story involves causal relations (the treats eaten by the animals causing red dots on their faces to disappear) that are plausible given children's naïve theories of causes of biological outcomes. Notaro, Gelman and Zimmerman (2001) found that children have domain boundaries in their causal reasoning. In line with that finding, Schulz et al. (2007) observed that pre-school children were less able to make use of current evidence when the evidence crossed boundaries of their naïve theories.

In constructing our story, we aimed for the candidate causal relations to be plausible to the children. The children's sensitivity to the observed evidence suggests that our choice of materials was appropriate. Equally important, because plausibility plays an identical role across all decomposition-function models we tested, the interpretation of our findings is not affected by children's prior domain knowledge of biological outcomes.

5.4 | Study 2

There are alternative explanations for why the children selected leaves in Study 1. The children's attention could be biased toward the second treat, the newer one. The children might simply have a bias toward leaves. Or they might have used a non-normative heuristic: pick the treat uniquely associated with the *fewest* animals with red dots after the intervention. Previous related experiments have not ruled out analogous hypotheses (Liljeholm & Cheng, 2007; Liljeholm et al., 2007). To rule out all three alternative explanations, Study 2 presented the same story but with the event frequencies in Table 2, to a separate group of pre-schoolers. As should be clear from the table, the heuristics and biases still predict choosing leaves. For example, as before, fewer animals had red dots after the intervention at the zoo than at the farm (one and two, respectively). The noisy-AND-NOT function, however, predicts choosing grain this time: the

"treatment" maintained the *same* preventive strength of 3/4 at the farm and the zoo—the leaves must therefore have no effect on red dots. The linear function now predicts (counterintuitively) that the leaves cause red dots. However, its recommended action coincides with the noisy-AND-NOT function's.

5.4.1Method

5.4.1.1 | Participants

The participants were 28 pre-school-age children (14 female) with a mean age of 4.38 years (range: 2.61–5.18 years, *SD* = 0.66 years). An additional two children began the task but were excluded for failure to attend to the story. Children were recruited similarly using the same criteria as for Study 1.

Table 2. *Event frequencies for Study 2*

	Farm	**Zoo**
Intervention	grain only	grain & leaves
Pre intervention: animals with dots	8/8	4/8
Post intervention: animals with dots	2/8	1/8
Number Cured	6	3
Fraction Cured	6/8	3/4

5.4.1.2 | Procedure

The procedure replicated that in Study 1 except that there were 16 trials in total, with the event frequencies for the farm and zoo animals as specified in Table 2. As in Study 1, to control for primacy and recency effects across contexts, the first trial at the farm and the zoo showed the same event type (an animal showing red dots before treatment but *no* red dots after treatment); likewise, the last trial at the two locations showed the same event type (an animal showing red dots before treatment but *no* red dots after treatment).

5.4.2 | Results and discussion

As before, the critical result concerned which treat children selected to make the animals' red dots go away. Fig. 5 shows that children's pattern of responses reversed in Study 2: contradicting the hypothesis that children responded based on one or more of the heuristics and biases, children were now significantly more likely to select the grain treat, X^2 (1, N = 28) = 5.14, p = .02.

The reversal of choices across Studies 1 and 2 was not due to age, even though the children in Study 2 were older on average. The same reversal was observed when mean age was controlled: for the children from the two studies ranging from 39 to 56 months, whereas three chose grain

and 11 chose leaves in Study 1, 10 chose grain and four chose leaves in Study 2, X^2 (1, N = 28) = 7.04, p = .008, for a test of a difference between experiments.

The opposite predominant choices across the two studies provide support for the noisy-AND-NOT function representing causal invariance, but cannot be explained by children applying a linear integration function or a non-normative heuristic, or had an attentional or other bias. Our results suggest that children (implicitly) distinguish between analytic and empirical knowledge, a distinction that enables them to assess the deviation of the empirical observations from *analytic* causal invariance across contexts.

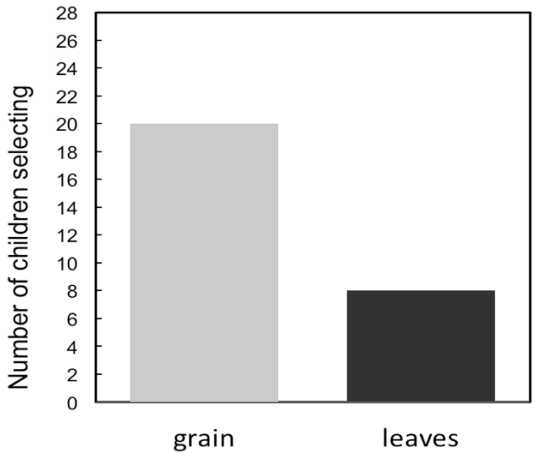

Fig. 5: Results from Study 2 depicting the number of children selecting the grain treat versus the leaves treat.

We replicated the pattern of results in Studies 1 and 2 in a variant in which the children were randomly assigned to the two studies, and the order of trials in each context (farm and zoo) was randomized for each child.

We note, however, that as with any pattern of observations, more complicated alternative hypotheses remain to be ruled out. For example, it is possible that children's opposite choices are instead due to the use of the linear function in combination with a bias toward the candidate with the more frequent pairing, the second treat, or leaves, with a shift in the weighing of the contributing factors adjusted to account for the observed change across studies.

5.5 | Replications of Studies 1 and 2: Random assignment across studies

We conducted a variant of our studies with another sample of 41 pre-schoolers (23 girls). The method was as before, except that the children were randomly assigned to our two studies, the story was presented on a laptop computer, and the order of the trials in each context (farm and zoo) was randomized for each child. The children's mean age was 4.05 years (SD = 0.76 years, range: 2.87–5.50 years). One additional girl in Replication Study 1 was excluded due to failure

to complete the task; three additional children in Replication Study 2 were excluded, one girl and one boy due to failure to complete the task and one boy due to experimenter error.

The same pattern of results as in Studies 1 and 2 was obtained. Whereas most of the children (15 out of 20) chose leaves in Replication Study 1, few of them (6 out of 21) chose leaves in Replication Study 2. The difference in choice between the replication studies was highly reliable, X^2 (1, N = 41) = 8.84, p = .0029. In each study, the children's choice was unlikely to be due to chance, X^2 (1, N = 20) = 5.0, p = .025 for Replication Study 1, and X^2 (1, N = 21) = 3.86, p < .05 for Replication Study 2.

5.6 | Adult analog of Studies 1 and 2

We replicated our two studies with 53 adults, undergraduate students at UCLA participating as an option to partially fulfill a course requirement. No participant was excluded. We followed the same method as in our pre-schooler replication, randomly assigning participants across studies, and minimally adapting it as follows. The materials were presented on a desktop computer, the participants were told that the materials came from an experiment conducted on pre-school children, and they read the story on their own. On the critical test trial, the participants were asked which treat would be their "best bet" for getting rid of the animals' red dots.

The pre-schoolers' pattern of results replicated: whereas most of the adult participants (21 out of 27) receiving the materials in Study 1 chose leaves, few of them (6 out of 26) receiving the materials in Study 2 chose leaves. The difference between the two groups was highly reliable, X^2(1, N = 53) = 15.86, p < .001.

Comparing our adult results with our pre-schooler results (pooling across pre-schoolers in our original experiments and in our replication), the performances of the two age groups were highly similar. Their responses did not statistically differ from each other, X^2(1, N = 76) = .03, p = .85 for Study 1 and X^2(1, N = 75) = .26, p = .61 for Study 2.

Our experimental materials and data were deposited at https://osf.io/v94jx/.

6 | General discussion

The goal of our present paper is to provide support for the essentiality of the concept of causal invariance, as a constraint during learning and as a criterion for belief revision during generalization, for the construction of useable causal knowledge. Our work aims to address real-life situations in which background causes may occur, possibly with different probabilities across contexts, where the set of possible causal representations is inherently too large to exhaustively evaluate. In support of our causal-invariance hypothesis, our results favor young children's spontaneous use of a causal-invariance function over the simpler linear function or more complex generalized linear function. Children's pattern of response suggests that they have analytic knowledge of how causes would combine their effects if each cause acts as if other causes were not there. Our results cannot be explained by a bias toward a particular candidate

cause or by a heuristic to choose the candidate less frequently paired with the undesired outcome.

Although there have been studies in the developmental literature testing causal Bayesian models (e.g., Lucas et al. 2014, Schulz et al., 2007) and the causal Markov assumption basic to causal Bayes nets (e.g., Gopnik et al., 2004; Sobel et al., 2004), no previous study has tested children's adoption of a causal framework that requires the causal-invariance assumption. Our findings suggest that the human causal-induction process was shaped to be capable of constructing useable causal knowledge in a reality that is accessible only via representations. Robust and wide-ranging evidence indicates that adults violate the causal Markov assumption (e.g., see Rottman & Hastie, 2014, 2016), even for the simplest causal networks involving three nodes (in comparison, Study 1 involves five nodes: grain, leaves, farm, zoo, and red_dots). What explains this violation, whereas support for adults' use of the causal-invariance assumption is strong (e.g., Buehner et al., 2003; Liljeholm & Cheng, 2007; Lu et al., 2008) and further strengthened by our studies on pre-schoolers? Our conjecture is that causal Bayes nets and human causal induction solve different problems (Park et al., 2022). Unlike generic causal Bayes nets in artificial intelligence (e.g., Pearl, 2000; Spirtes, Glymour, & Scheines, 1993/2000), natural human causal induction did not evolve in response to situations in which data input are encoded in predefined variables supplied by "users." The representation of variables that serve our purposes is ours alone to formulate and reformulate. This difference in the input to the causal-induction process defines two distinct computational-level problems (Marr, 1982). Human causal induction solves the representation-formulation problem that causal Bayes nets cannot solve and does not require the Markov assumption used in artificial intelligence for pruning networks encoded in predefined variables.

One of the reasons we designed our materials to involve a context change is that it enables both quantitatively and qualitatively different predictions according to causal and associative models, the latter including the widely used and supposedly normative logistic regression model. It is possible that the children failed to attend to the context change so that they treated all the animals as being of the same type from the same place.[7] If that was indeed the case, the qualitatively different predictions would not apply. Although we cannot empirically state that children in our current studies did attend to the context change, we have several reasons to expect that they did. First, we made sure to emphasize the shifting context both visually and verbally within the story and trained the experimenters to be emphatic in conveying that change in the telling of the story to the children. Second, in other studies with pre-school-age children, even incidental and irrelevant changes, such as the background color of an image in a word-learning task, result in large effects on both memory and generalization of the words (Vlach & Sandhofer, 2011). Third, what comprises the "same context" broadens over development, such that infants and young children may be more sensitive to context changes than adults (Rovee-Collier & Cuevas, 2009; also see Sandhofer & Schonberg, 2020). Context changes that are salient to adults may also be unlikely to be unnoticed by children.

Although the no-context-change hypothesis seems implausible in our view given our methodology and previous findings, it would nonetheless be interesting to test, to provide evidence for or against the children taking context change into account. To test the no-context-

change hypothesis, future work might consider creating four conditions, two that are identical to Studies 1 and 2, respectively (the control groups) and two that differ from our studies only in that there is no context change; for example, all the animals are farm animals from the brothers' farm (the no-context-change groups). If children in our studies in fact ignored context, the no-context-change groups should respond similarly as the respective control groups. However, children in the no-context-change conditions may have greater difficulty basing their treat choice on the entire sequence of trials in the conditions (40 in the Study 1 analog), rather than making an inference on the basis of half of those trials in one context at a time as in our studies. The greater difficulty may lead to picking a treat randomly or using a heuristic to answer the question. A difference in treat choice between the no-context-change groups and their respective controls would reject the no-context-change null hypothesis.

In the rest of this section, we first explain how our work relates to recent work on actual-causation judgments in the cognitive science literature. We then explain why, for any particular variable type (e.g., binary with a present and absent value), there cannot be multiple characterizations of the *sameness* of causal influence that justifies generalization across contexts.

6.1 | How our work relates to work on explanatory scope and counterfactual explanation of actual causation

Our present paper on pre-schoolers' intuitive reasoning shares the treatment of causal invariance during learning as an aspiration with work on the history of science (Kuhn, 1962/2012; Woodward, 2000). And it is complementary to work on the role of causal invariance during the application of previously learned causal knowledge (e.g., Blanchard, Vasilyeva, & Lombrozo, 2018; Danks, 2013; Hitchcock, 2012; Icard, Kominsky, & Knobe, 2017; Kominsky, Phillips, Gerstenberg, Lagnado, & Knobe, 2015; Lombrozo, 2010; Nagel & Stephan, 2015, 2016; Stephan & Waldmann, 2016, 2018; Woodward, 2006, 2021). This line of work has concerned actual causation (i.e., causes of single events) rather than "type" causation. Although there are common issues that underlie causal judgments for actual and type causation (e.g., Blanchard et al., 2018; Danks, 2013; Lombrozo, 2010; Nagel & Stephan, 2016; Woodward, 2006), neither kind of causal judgment carries implications for causal learning and knowledge revision involving type causation. In contrast, our line of work has focused on such implications (Carroll & Cheng, 2010; Cheng et al., 2013; Cheng & Lu, 2017).

Some of the work on actual causation explains why in some cases causal explanations with a broader explanatory scope are more acceptable, whereas in other cases those with a narrower explanatory scope are more acceptable (e.g., Blanchard et al., 2018; Woodward, 2006; cf. Lien & Cheng, 2000). Others explain why counterfactual conditions that would have made a difference to an actual outcome do not all point to equally acceptable actual causes of the outcome. For example, counterfactual dependence that concerns a prevalent or normal condition indicates an actual cause, whereas one that concerns a rare or abnormal condition does not (Hitchcock, 2012; Lombrozo, 2010; Nagel & Stephan, 2015). Yet other work shows that the perceived goodness of an explanation depends on the causal strength of the causal type (Nagel & Stephan, 2016) or on an interaction between moral norms and the causal structure of the

situation (Icard et al., 2017; Kominsky et al., 2015). As this brief summary shows, this diverse line of work shows a preference for causal invariance in a far more varied sense than ours. Nonetheless, these counterfactual accounts of causal judgment make an invariance assumption analogous to ours in that they require that individual causes have the same influence in the counterfactual possibilities considered as they did in the actual event.[8]

6.2 | Why, for any given outcome-variable type, there cannot be multiple characterizations of the sameness of the operation of a causal mechanism

It is tempting to regard the accuracy of a model's prediction as a gauge of the model's distance from "the truth." However, models are representations, which are inevitably based on assumptions. If a causal model assumes that a target cause's influence on the outcome *interacts* with (potentially unknown) causes in the learning context (e.g., combining the influences of the grain treat and background causes at the farm according to the linear function), then that causal model's accurate prediction of the outcome in a novel context (the grain predicting the red-dot outcome perfectly at the zoo under the linear function) gives support to the interaction (i.e., the grain acts differently at the zoo and the farm). Thus, instead of supporting that the target cause operates in the same manner in a novel context, accurate prediction under an interaction assumption supports the opposite (i.e., the grain will act differently in another context). There are infinitely many ways of acting differently in contexts that differ in different ways. An accurate prediction by such a causal model would be a matter of coincidence (which our experimental design contrived) rather than a revelation of a generalizable causal structure.

One might counter: Are there not multiple legitimate definitions of the "sameness of causal capacity" for an outcome-variable type? To answer this question, we note our assumption that we live in a world in which there is only one way for a causal mechanism to operate *unchanged* across contexts, whatever that mode of operation is. This question is different from the issue of whether two objects similar and dissimilar on various dimensions are in the same category (Goodman, 1972; Medin, 1989). There are many ways for objects, which necessarily have multiple properties, to be considered the same. Instead, our question here concerns *sameness along one single dimension in question*, for example, the sameness of the number of photons cast by a lamp per unit time on an area when another lamp shines on it and when no other lamp shines on it. If two lamps each cast 10 lumens of light in an area, and they do not change their illuminance in the context of the other lamp, together they must cast 20 lumens of light in the area. Although different variable types have different integration functions characterizing "sameness of causal capacity" (e.g., linear for continuous variables, noisy-OR for binary variables with a present value and an absent value, wave addition for waves), for any given variable type, there is one and only one way to coherently define that sameness of operation in the world for the purpose of generalization across contexts.

Below we illustrate the uniqueness of the definition of *sameness* of causal capacity across contexts for binary variables with a "present" and an "absent" value by way of a negative example. Note that the relative frequency of an outcome being present in a set of entities is a proportion involving cardinal numbers (e.g., three out of 12 patients), which are on a ratio scale.

Units on this scale (e.g., 0 patients, 3 patients) have a physical meaning and are not arbitrary.[9]Suppose a pill "invariably" doubles the probability of a desirable binary outcome in patients across two contexts, where the baseline proportion of patients with the outcome before treatment is 1/4 and 1/2, respectively. Although "invariably doubling" the probability of an outcome may linguistically suggest an invariant capacity of the drug, the underlying causal mechanism in fact cannot physically be operating the same way in each patient across these contexts. For each context, consider the question: what must the pill do in *patients who do not already show the outcome before treatment*? This is the crucial subgroup of patients to focus on because they are the only ones for whom the drug potentially changes the binary outcome (from absent to present). Thus, in the context where 1/4 of the patients before treatment show the outcome, focus on the remaining 3/4 of the patients. Doubling the probability of the outcome from the baseline of 1/4 requires that the drug has a 1/3 probability of producing the outcome in each patient: in order to give an overall proportion of 1/2 after treatment, 1/3 of the 3/4 who do not already show the outcome must now show the outcome. But, in the context where 1/2 of the patients before treatment show the outcome, doubling the probability of the outcome requires that, in the remaining half who do not already show the outcome, the drug in each patient *always* produces the outcome. And, of course, if the baseline probability of the outcome is 0, doubling that probability means the drug has no effect. Thus, the drug that "invariably doubles" the probability of the outcome in fact varies the full range from being noncausal in one context to being a deterministic cause in another context.

To fit the linguistic description of "invariably doubling" the probability of a binary outcome, the pill in each patient in the crucial subgroup would have to adapt its influence *depending on* the proportion of the complementary subgroup, namely, patients who already show the outcome before treatment. A magnitude of influence that depends on "knowing" the baseline probability (i.e., is a function of the influence of the contextual causes) violates the "sameness of causal capacity" across contexts, by definition. In contrast, the noisy-logical integration functions, and only those functions, defining invariant causal capacity as they do according to the product definition of independence in probability theory (e.g., Feller, 1968; Jaynes, 2003), by definition hold the causal capacity of the drug constant regardless of the contextual causal influences.

In conclusion, in the infinite hypothesis space of possible causal representations, prediction error defined in terms of deviation from the expected outcome assuming causal invariance among hypothesized whole causes is not a mere wish or a convenience, but a rational navigation device for constructing useable causal knowledge. Our findings indicating the early use of a probabilistic causal-invariance function—embodying the aspiration of finding *forces of change that themselves remain unchanged*—suggest that the invariance of causal knowledge across contexts, along with parsimony and logical consistency, is an essential cognitive constraint.

Appendix

Here we present model predictions for the inferential problems used as materials in Studies 1 and 2. We first present predictions according to both Bayesian and frequentist variants of associative models (Fienberg, 2007; Griffiths & Tenenbaum, 2005; Tenenbaum & Griffiths, 2001; Wickens, 1989; Yuille & Lu, 2007), including predictions according to the logistic regression model (e.g., McCullagh & Nelder, 1989; Wickens, 1989). We classify a model as *associative* if it does not allow an evaluation of the simplest causal model for the data. We then present predictions based on causal (i.e., non-associative) Bayesian models (Yuille & Lu, 2007), including both structure-learning and parameter-estimation models. We apply these models to estimate the causal strengths or evaluate the causal status of the farm intervention (the grain alone), the zoo intervention (the grain and the leaves), and the leaves alone. For brevity, in the calculations below we denote the two interventions by "farm_iv" and "zoo_iv," respectively, and the outcome in question, "red dots on the face," by "red." To relate to models in the psychological literature (e.g., Cheng, 1997; Jenkins & Ward, 1965; Pearce, 1987, 1994; Rescorla & Wagner, 1972), our analysis partitions all causes of *e* into two *causal paths*. One path is represented by candidate cause *c* and the chain of variables connecting *c* to *e* that are changed or potentially changed, by manipulating *c*. The other path is represented by alternative causes *a*, where *a* is a composite of causes of *e* in the context, that is, causes that are not part of the causal path from *c* to *e*(these include enabling conditions of *e*). When *c* is absent, the occurrence of *e* is explained by *a*.

Predictions of the associative view for Study 1: Parameter-estimation and structure-learning models

Causal strengths of the treats according to the linear model

When multiple causes are present, the occurrence of *e* is explained by a sum of the associative strengths (i.e., weight) of the causes. Both Bayesian and frequentist models can adopt the linearity assumption (e.g., Griffiths & Tenenbaum, 2005; Tenenbaum & Griffiths, 2001; Yuille & Lu, 2007). Let w_c represent the associative strength between *c* and *e*, and w_a represent that between *a* and *e*. According to the linearity assumption in this model, when there is *no confounding* (i.e., *a* occurs just as often whether or not *c* occurs), Eqs. A1 and A2 follow:

$$P(e = 1 \mid c = 0) = w_a. \qquad \text{(Eq. A1)}$$

$$P(e = 1 \mid c = 1) = w_c + w_a. \qquad \text{(Eq. A2)}$$

According to this model, for Study 1 (see data pattern in Table 1),

$$w_{grain} = w_{farm\_iv} = P(red = 1|farm\_iv = 1) - P(red = 1|farm_{iv} = 0) = \frac{6}{10} - \frac{9}{10} = -3/10 \qquad \text{(Eq. A3)}$$

w_{farm_iv} is equal to w_{grain} because grain is the only intervention at the farm. Likewise,

$$w_{zoo\_iv} = P(red = 1|zoo\_iv = 1) - P(red = 1|zoo_{iv} = 0) = \frac{1}{10} - \frac{4}{10} = -\frac{3}{10} \qquad \text{(Eq. A4)}$$

Because grain and leaves jointly are the intervention at the zoo, the linear assumption implies:

$$w_{zoo\_iv} = w_{grain} + w_{leaves} \quad \text{(Eq. A5)}$$

Therefore, by Equations A3, A4, and A5:

$$w_{leaves} = 0.$$

These weight estimates for grain and leaves correspond to the Bayesian maximum-likelihood estimates of causal strengths assuming the linear generating function.

Causal strengths of the treats and the causal status of leaves according to logistic regression

Logistic regression is a *generalized linear model* (GLM; Fienberg, 2007; McCullagh & Nelder, 1989, to be distinguished from a "general linear model") for predicting the probability of the occurrence of a binary outcome (e.g., red dots vs. no red dots in our farm-and-zoo problem) by fitting data to a logistic function of a linear combination of predictor variables (e.g., grain, leaves, contextual causes). A logistic model with n predictor variables $x_1, x_2, \ldots, x_n$ states

$$ln\left[\frac{P(outcome)}{1-P(outcome)}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n, \quad \text{(Eq. A6)}$$

where P(outcome) is the probability of the binary outcome.

The statistical results we report here are from a logistic regression analysis of the pre-intervention and post-intervention event frequencies in Table 1, with each event frequency multiplied by 3: Out of 30 farm animals, who were all given the grain treat, 27 had red dots before the treat, and 18 had red dots after the treat; out of 30 zoo animals, who were all given the grain and the leaves, 12 had red dots before the treats, and three had red dots after the treats. We used a larger sample size to allow a clearer interpretation, so that failure to reject the null hypothesis that the leaves treat has a weight of 0 is clearly not due to the sample size being small. (Children, like adults, are likely to draw causal conclusions based on the pattern of events, mostly ignoring sample size; see Lu et al., 2008). The weights, w_{farm} and w_{zoo}, are the weights for the farm and zoo variables and are different from the weights for "farm_iv" and "zoo_iv", the *interventions* at the respective locations.

Our analysis shows the following weights:

$w_{farm} = 2.20$ [Wald χ^2 (1) = 13.0, $p < .001$];

$w_{grain} = -1.80$ [Wald χ^2 (1) = 6.30, $p = .01$];

$w_{zoo} = -0.41$ [Wald χ^2 (1) = 1.18, $p = .28$];

$w_{leaves} = 0$ [Wald χ^2 (1) = 0, $p = 1$; that is, w_{leaves} is as far from being different from 0 as is possible]. Our analysis treats the two contexts (farm and zoo) as a fixed effect, consistent with our intended interpretation of the scenario.

If the context is erroneously treated as a random effect in the analysis, the same recommendation to give the new animals the grain treat is reached: $w_{grain} = -1.50$, $p = .03$ and $w_{leaves} = -0.34$, $p = .73$. The leaves' weight in that case is nonetheless small and insignificant.

Causal status of the leaves treat according to a Bayesian model with the linear generating function

Like logistic regression, a linear variant of a Bayesian structure-learning model (Yuille & Lu, 2007) also prescribes choosing grain. This model is the Bayesian analog of an associative frequentist test applied to our farm-and-zoo scenario, developed for an analog of our farm-and-zoo problem that asks a variant of our critical question: whether the second treat, leaves, is a preventer of red dots. The model computes *causal support* for the leaves treat: the log-likelihood ratio between two causal graphs in which the leaves treat, respectively, does and does not prevent red dots. A positive causal-support value means the evidence is in support of the leaves treat being a preventive cause of red dots; a negative value means the evidence is against that hypothesis; a value of 0 indicates neutrality. For the data pattern in Table 1, causal support for the leaves treat under the linear generating function is −0.92, indicating evidence *against* leaves preventing red dots. In summary, all associative models—both the parameter-estimation and structure-learning variants of the linear Bayesian model and the logistic regression model—prescribe choosing grain.

Predictions of the causal invariance view for Study 1 according to Bayesian models

To illustrate the causal-invariance view, we use *causal power* (Cheng, 1997; Sheps, 1958; "power" in the sense of capacity, Cartwright, 1989) and a Bayesian structure-learning model adopting the *noisy-logical* generating functions (Yuille & Lu, 2007). These models assume causal invariance.[10] Causal power is the Bayesian maximum-likelihood estimator of the strength of the influence of *c* on *e* assuming that cause *c* and the background cause in the learning context combine their strengths in a noisy-logical form and that the two causes of *e* both have a uniform prior of causal strength (Tenenbaum & Griffiths, 2001).

Applying the causal-invariance assumption here means that the grain operates with the same causal mechanism across the farm and zoo contexts, which implies that *for every animal* (all 20), grain has the *same* capacity to remove red dots. We denote the power of a candidate *c* to prevent or remove red dots by p_c. Applying Eq. 3 to the scenario,

$$p_{farm\_iv} = p_{grain} = \frac{9/10 - 6/10}{9/10} = 1/3\,. \qquad \text{(Eq. A7)}$$

Likewise,

$$p_{zoo\_iv} = \frac{4/10 - 1/10}{4/10} = 3/4\,. \qquad \text{(Eq. A8)}$$

When red dots are removed at the zoo, in the presence of both grain and leaves, they are removed by grain or by leaves. The causal invariance of grain and leaves (by Eq. 1 in main manuscript) implies:

$$p_{zoo\_iv} = p_{grain} + p_{leaves} - p_{grain} \cdot p_{leaves}. \qquad \text{(Eq. A9)}$$

From Eqs. A7, A8, and A9, it follows that $p_{leaves} = 5/8$. Because 5/8 is greater than 1/3 (i.e., the leaves treat brings stronger relief than grain), the causal view prescribes choosing leaves. The same answer follows from Equation 2 (in main manuscript) assuming the causal invariance of grain across contexts.

According to Yuille and Lu's (2007) structure-learning noisy-logical model, the causal-support value is 1.21, indicating that leaves prevent red dots. In summary, for Study 1, all causal-invariance models -- both parameter-estimation and structure-learning variants -- prescribe choosing leaves.

Predictions for Study 2: Bayesian parameter-estimation and structure-learning models

Analogous calculations for Study 2 (see data pattern in Table 2) yield the following results. For the linear model, we obtain $w_{grain} = -3/4$ and $w_{leaves} = +3/8$. The positive value for leaves means that leaves cause red dots (rather than remove them). Therefore, the model recommends choosing grain to give the new animals to remove the red dots. Logistic regression is not applicable to this scenario because one of the probabilities of red dots is 1. For the linear causal-support model (Yuille & Lu, 2007), the support value for leaves is 0.05, indicating weak evidence for leaves preventing red dots.

Causal invariance also predicts choosing the grain over the leaves treat, but in contrast to the linear models, it predicts that the leaves treat has no influence on red dots, in accord with intuition. This yields the following preventive strengths: $p_{\text{grain}} = 3/4$ and $p_{\text{leaves}} = 0$. Consistent with the maximum-likelihood causal strength estimates, the noisy-logical causal-support model gives a support value of −0.16 for leaves, indicating no support for the hypothesis that leaves prevent red dots.

Acknowledgments

The research reported in this article was supported by AFOSR FA 9550-08-1-0489. We thank Joseph Burling, Chris Carroll, Clark Glymour, Reid Hastie, Hongjing Lu, Tom Wickens, and Yingnian Wu for helpful discussions, Jonathan Kominsky and Tamir Kushnir for helpful comments, and Maryann Francis, Natalie Peri, Aaron Placencia, and Michael Venditti for conducting the experiments. Last but by no means least, we thank Om Bleicher for awesome farm-and-zoo storybook artwork. Parts of this article have appeared in the *Proceedings of the 35th Annual Conference of the Cognitive Science Society* and the *Proceedings of the 39th Annual Conference of the Cognitive Science Society*.

Notes

[1] A more general form of the concept of deterministic causal invariance.
[2] To focus on causal rather than statistical inference, we assume a reasonable sample size and very narrow confidence intervals in both experiments.
[3] In other words, whereas analytic knowledge is justified by the validity of the inference, empirical knowledge is justified by correspondence to observations/"truth" in the world. (See Skyrms, 2000, for the related difference between deduction and induction.) Thus, physics is an empirical science. And for our example conditional proposition to be valid, person *x* need not be a surgeon. In fact, the relation between surgeons and medical doctors would still be valid if there happens to be no surgeons in the world.
[4] See Ichien and Cheng (2022), for dataset examples illustrating this logical implication.
[5] Our present paper does not address *how* causal knowledge is revised to restore causal invariance. The nature of the revision would depend on the domain and the causal reasoners' knowledge and imagination.
[6] We confirmed this conclusion based on Fig. 1 with a logistic regression analysis, treating the two contexts (farm and zoo) as fixed effects, consistent with our intended interpretation of the scenario. The same conclusion results, however, if the contexts are erroneously treated as random effects. (See prediction details in our Appendix.)
[7] We thank Jonathan Kominsky for suggesting this possibility.
[8] Thanks to Jonathan Kominsky for noting this commonality.
[9] Transformations of the scale would lose that meaning.
[10] "Noisy-logical" is a general label for independent probabilistic generative and preventive influence of causes and compositions of them (Yuille & Lu, 2007). The independence assumption can be weakened under the following condition without affecting the predictions. Suppose unobserved background causes (e.g., oxygen) interact with the candidate (striking a match), so that independent influence (on fire) does not hold (see Cheng, 2000). Even in that case, if the interacting background factors occur with the same probability in the learning and application contexts, the candidate cause would produce the outcome the same way across contexts, with the same causal power, as if there is no interaction. The same predictions as presented here would therefore follow.

Conflict of Interest

The authors have no known competing financial interests or personal relationships that could have influenced the work reported in this article.

References

Beckers, T., De Houwer, J., Pineño, O., & Miller, R. R. (2005). Outcome additivity and outcome maximality influence cue competition in human causal learning. Journal of *Experimental Psychology: Learning, Memory, & Cognition, 31*, 238-249.

Beckers, T., Miller, R. R., De Houwer, J., & Urushihara, K. (2006). Reasoning rats: Forward blocking in Pavlovian animal conditioning is sensitive to constraints of causal inference. *Journal of Experimental Psychology*: *General*, *135*, 92-102.

Blanchard, T., Vasilyeva, N., & Lombrozo, T. (2018). Stability, breadth and guidance. *Philosophical Studies*, *175*, 2263–2283.

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, 301-338.

Buehner, M., Cheng, P.W., Clifford, D. (2003) From covariation to causation: A test of the assumption of causal power. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, pp. 1119-1140.

Carroll, C.D. & Cheng, P.W. (2010) The induction of hidden causes: Causal mediation and violations of independent causal influence. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, eds S. Ohlsson & R. Catrambone (pp. 913-918). Austin, TX: Cognitive Science Society.

Challenges in Reproducible Research (2016). *Nature*: http://www.nature.com/nature/focus/reproducibility/index.html - perspectives

Cheng, P.W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*, 367-405.

Cheng, P.W. (2000). Causality in the mind: Estimating contextual and conjunctive causal power. In F. Keil & R. Wilson (Eds.), *Explanation and cognition* (pp. 227-253). Cambridge, MA: MIT Press.

Cheng, P.W., Liljeholm, M. & Sandhofer, C. (2013). Logical consistency and objectivity in causal learning. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 2034-2039). Austin, TX: Cognitive Science Society.

Cheng, P.W., Liljeholm, M., Sandhofer, C.M. (2017). Analytic causal knowledge for constructing useable empirical causal knowledge: Two experiments on preschoolers. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Cheng, P.W. & Lu, H. (2017). Causal invariance as an essential constraint for creating a causal representation of the world: Generalizing the invariance of causal power. In M.R. Waldmann (Ed). *The Oxford Handbook of Causal Reasoning.* Oxford, England: Oxford University Press.

Cheng, P.W., Novick, L.R., Liljeholm, M. & Ford, C. (2007). In M. O'Rourke (Ed.), *Topics in contemporary philosophy*, Volume *4*: *Causation and explanation* (pp. 1 – 32). Cambridge, MA: MIT Press.

Cook, C. et al. (2011) Where science starts: spontaneous experiments in preschoolers' exploratory play. *Cognition*, *120*, 341–349.

Danks, D. (2013). Functions and cognitive bases for the concept of actual causation. *Erkenntnis*, *78*, 111-128.

Denison, S., Xu, F. (2010). Twelve- to 14-month-old infants can predict single-event probability with large set sizes, *Developmental Science, 13*, 798-803.

De Houwer, J., Beckers, T., & Vandorpe, S. (2005). Evidence for the role of higher-order reasoning processes in cue competition and other learning phenomena. *Learning & Behavior*, *33*, 239–249.

Feller, W. (1968). *An introduction to probability theory and its applications* (3rd edition). New York: Wiley.

Fienberg, S.E. (1980/2007). *The analysis of cross-classified categorical data* (2^{nd} edition). Cambridge: MIT Press.

Goodman, N. (1972). "Seven strictures on similarity." In N. Goodman (Ed.), *Problems and projects*. New York: Bobbs-Merrill.

Gopnik, A. (2009). *The philosophical baby*. New York: Farrar, Straus and Giroux.

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*, 1-30.

Griffiths, T.L., & Tenenbaum, J.B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51*, 285-386.
Gweon, H. & Schulz, L.E. (2011) 16-month-olds rationally infer causes of failed actions. *Science*, *332*, 1524.
Hitchcock, C. (2012). Portable causal dependence: A tale of consilience. *Philosophy of Science*, *79*, 942-951.
Hume, D. (1739/1987) *A treatise of human nature* (2nd edition, Clarendon Press, Oxford).
Hume D. (1777/1975) *An enquiry concerning human understanding and concerning the principles of morals*, eds Selby-Bigge LA & Nidditch PH (3rd edition). Oxford: Clarendon Press.
Icard, T.F., Kominsky, J.F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition*, *161*, 80-93.
Ichien, N. & Cheng, P.W. (2022). Revisiting Hume in the 21st century: The possibility of generalizable causal beliefs given inherently unobservable causal relations. In A. Wiegmann & P. Willemsen (Eds.), *Advances in Experimental Philosophy of Causation*. (pp. 7 – 34) London, UK: Bloomsbury Press.
Ioannidis, J.P.A. (2005). Why most published scientific findings are false. *PLoS Medicine*, *2*(*8*), e124.
Jaynes, E.T. (2003) *Probability theory: The logic of science*. Cambridge: Cambridge University Press.
Jenkins, H.M., Ward, W.C. (1965). Judgment of contingency between responses and outcomes, *Psychological Monographs*: *General and Applied*, 79 (1, Whole No. *594*).
Kominsky, J.F., Phillips, J., Gerstenberg, T., Lagnado, D.A. & Knobe, J. (2015). Causal superseding. *Cognition*, *137*, 196-209.
Kuhn, T.S. (1962/2012). *The structure of scientific revolutions* (50th anniversary edition). Chicago: University of Chicago Press.
Kushnir, T., & Gopnik, A. (2005). Young children infer causal strength from probabilities and interventions. *Psychological Science*, *16*, 678-683.
Kushnir, T., & Gopnik, A. (2007). Conditional probability versus spatial contiguity in causal learning: Preschoolers use new contingency evidence to overcome prior spatial assumptions. *Developmental Psychology*, *44*, 186-196.
Lagnado, D.A., & Sloman, S. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory & Cognition, 30*, 856–876.
Lapidow, E. & Walker, C. M. (2021). Rethinking the "gap": Self-directed learning in cognitive development and scientific reasoning. *WIREs Cognitive Science*.
Legare, C.H., Gelman, S.A., & Wellman, H.M. (2010) Inconsistency with prior knowledge triggers children's causal explanatory reasoning. *Child Development*, *81*, 929–944.
Legare, C.H. (2012) Exploring exploration: explaining inconsistent information guides hypothesis-testing behavior in young children. *Child Development*, *83*, 173–185.
Lien, Y.W. & Cheng, P.W. (2000). Distinguishing genuine from spurious causes: A coherence hypothesis. *Cognitive Psychology*, *40*(2), 87–137.
Liljeholm, M., & Cheng, P.W. (2007). When is a cause the "same"? Coherent generalization across contexts. *Psychological Science*, *18*, 1014-1021.
Liljeholm, M., Cheng, P.W. & Leung, B. (2007). Revision of simple causal hypotheses: Inferring interaction across multiple contexts. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, eds McNamara DS & Trafton JG (pp. 1223 – 1228). Austin, TX: Cognitive Science Society.
Lombrozo, T. (2010). Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, *61*, 303-332.
Lovibond, P.F. (2003). Causal beliefs and conditioned responses: Retrospective revaluation induced by experience and by instruction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 97–106.

Lovibond, P.F., Been, S-L., Mitchell, C.J., Bouton, M.E., & Frohardt, R. (2003). Forward and backward blocking of causal judgement is enhanced by additivity of effect magnitude. *Memory & Cognition, 31*, 133–142.

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

Lu, H., Rojas, R. R., Beckers, T., & Yuille, A. L. (2016). A Bayesian theory of sequential causal learning and abstract transfer. *Cognitive Science, 40*, 404–439.

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

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

Marr, D. (1982). *Vision*. New York: Freeman.

McCullagh, P. & Nelder, J.A. (1989). *Generalized linear models* (2nd edition, CRC Press, Boca Raton).

Medin, D.L. (1989). Concepts and conceptual structure. *American Psychologist, 44,* 1469–1481.

Mehta, R. & Williams, D.A. (2002). Elemental and configural processing of novel cues in deterministic and probabilistic tasks. *Learning and Motivation*, *33*, 456–484.

Melchers, K.G., Lachnit, H., & Shanks, D.R. (2004). Past experience influences the processing of stimulus compounds in human Pavlovian conditioning. *Learning and Motivation*, *35*(3), 167-188.

Nagel, J., & Stephan, S. (2015). Mediators or alternative explanations: Transitivity in human-mediated causal chains. In D. C. Noelle et al. (Eds.) *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 1691– 1697). Austin, TX: Cognitive Science Society.

Nagel, J. & Stephan, S. (2016). Explanations in causal chains: Selecting distal causes requires exportable mechanisms. In Papafragou, A., Grodner, D., Mirman, D., and Trueswell, J. C. (eds.) *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 806-811). Austin, TX: Cognitive Science Society.

NIH News Release (December 30, 2020). https://www.nih.gov/news-events/news-releases/peer-reviewed-report-moderna-covid-19-vaccine-publishes

Notaro, P.C., Gelman, S.A. & Zimmerman, M.A. (2001). Biases in reasoning about the consequences of psychogenic bodily reactions: Domain boundaries in cognitive development. *Merrill-Palmer Quarterly*, *48*, 427–449.

Novick, L.R. & Cheng, P.W. (2004). Assessing interactive causal influence. *Psychological Review, 111,* 455-485.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, *349*, 943: DOI: 10.1126/science.aac4716.

Park, J., McGillivray, S., Bye, J.K. & Cheng, P.W. (2022). Causal invariance as a tacit aspiration: Analytic knowledge of invariance functions. *Cognitive Psychology*. https://authors.elsevier.com/sd/article/S0010-0285(21)00055-4

Pearce, J.M. (1987). A model for stimulus generalization in Pavlovian conditioning. *Psychological Review*, *94*, 61-73.

Pearce, J.M. (1994). Similarity and discrimination: A selective review and a connectionist model. *Psychological Review*, *101*, 587-607.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Mateo, CA: Morgan Kaufmann.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, England: Cambridge University Press.

Rescorla, R.A. & Wagner, A.R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In *Classical conditioning II: Current theory and research*, eds. Black, A.H., Prokasy, W.F. (Appleton-Century Crofts, New York), pp. 64-99.

Rottman, B.M. & Hastie, R. (2014). Reasoning about causal relationships: Inferences on causal networks. *Psychological Bulletin, 140*, 109–139. http://dx.doi.org/10.1037/a0031903.

Rottman, B.M. & Hastie, R. (2016). Do people reason rationally about causally related events? Markov violations, weak inferences, and failures of explaining away. *Cognitive Psychology*, *87*, 88–134

Rovee-Collier, C., & Cuevas, K. (2009). Multiple memory systems are unnecessary to account for infant memory development: An ecological model. *Developmental Psychology*, *45*(1), 160–174. https://doi.org/10.1037/a0014538

Saffran, J.R., Aslin, R.N. & Newport, E.L. (1996). Statistical learning by 8-month-old infants. *Science, 174*: 1926-1928.

Sandhofer, C. M., & Schonberg, C., (2020). Multiple examples support children's word learning: The roles of aggregation, decontextualization, and memory dynamics. In J. Childers, S. Graham, & L. Namy (Eds.) *Learning Language and Concepts from Multiple Examples in Infancy and Childhood.* (pp 159-178). Springer. https://doi.org/10.1007/978-3-030-35594-4

Schulz, L. (2012). The origin of inquiry: inductive inference and exploration in early childhood. *Trends in Cognitive Science*, *16* (*7*), 382-389.

Schulz, L., Bonawitz, E. & Griffiths, T. (2007). Can being scared make your tummy ache? naive theories, ambiguous evidence and preschoolers' causal inferences. *Developmental Psychology*, *43*, 1124-1139.

Shanks, D. R. & Darby, R. J. (1998). Feature- and rule-based generalization in human associative learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *24*, 405–415.

Sheps, M.C. (1958). Shall we count the living or the dead? *The New England Journal of Medicine*, *259*: 1210-1214.

Skyrms, B. (2000). *Choice and Chance: An Introduction to Inductive Logic* (4th rev. ed.). Belmont: Wadsworth.

Sloman, S.A., & Lagnado, D. (2005). Do we ''do''? *Cognitive Science*, *29*, 5–39.

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*, 303-333.

Spellman, B. A. (1996). Acting as intuitive scientists: Contingency judgments are made while controlling for alternative potential causes. *Psychological Science, 7*(6), 337–342. https://doi.org/10.1111/j.1467-9280.1996.tb00385.x

Spirtes, P., Glymour, C. & Scheines, R. (1993/2000). *Causation, prediction and search* (2nd edition), MIT Press: Cambridge.

Stephan, S. & Waldmann, M.R. (2016). Answering causal queries about singular cases. In Papafragou, A., Grodner, D., Mirman, D., and Trueswell, J. C. (eds.) *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 2795-2801). Austin, TX: Cognitive Science Society.

Stephan, S. & Waldmann, M.R. (2018). Preemption in singular causation judgments: A computational model. *Topics in Cognitive Science*, 10, 242–257.

Tenenbaum, J.B. & Griffiths, T.L. (2001). Structure learning in human causal induction. In T.K. Leen, T.G. Dietterich, & V. Tresp (Eds.) *Advances in neural information processing systems*, *13* (pp. 59–65). Cambridge, MA: MIT Press.

Urcelay, G.P. & Miller, R.R. (2010). On the generality and limits of abstraction in rats and humans. *Animal Cognition, 13*, 21–32.

Vlach, H. A. & Sandhofer, C. (2011). Developmental Differences in Children's Context-Dependent Word Learning. *Journal of Experimental Child Psychology*, *108,* 394-401. doi:10.1016/j.jecp.2010.09.011

Waldmann, M.R. (2007). Combining versus analyzing multiple causes: How domain assumptions and task context affect integration rules. *Cognitive Science*, *31*, 233-256.
Wheeler, D.S., Beckers, T. & Miller, R.R. (2008). The effect of subadditive pretraining on blocking: Limits on generalization. *Learning &* Behavior, *36* (*4*), 341-351.
Wickens, T. D. (1989). *Multiway contingency tables analysis for the social sciences.* Erlbaum Associates: Hillsdale.
Woodward, J. (2000). Explanation and invariance in the special sciences. *British Journal of the Philosophy of Science*, *51*, 197-254.
Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation (Oxford Studies in the Philosophy of Science)*. Oxford, England: Oxford University Press.
Woodward, J. (2006) Sensitive and insensitive causation. *Philosophical Review*, *115*, 1–50.
Woodward, J. (2010). Causation in biology: stability, specificity, and the choice of levels of explanations. *Biological Philosophy*, *25*, 287–318. DOI 10.1007/s10539-010-9200-z
Woodward, J. (2021). *Causation with a human face: Normative theory and descriptive psychology*. Oxford, UK: Oxford University Press.
Wu, M. & Cheng, P.W. (1999). Why causation need not follow from statistical association: Boundary conditions for the evaluation of generative and preventive causal powers. *Psychological Science*, *10*, 92-97.
Yuille, A. L., & Lu, H. (2008). The noisy–logical distribution and its application to causal inference. In J. C. Platt, D. Koller, Y. Singer & S. Rowets (Eds.), *Advances in neural information processing systems*, *20* (pp. 449–456). Cambridge, MA: MIT Press.
Zimmer-Hart, C. L., & Rescorla, R. A. (1974). Extinction of Pavlovian conditioned inhibition. *Journal of Comparative and Physiological Psychology*, *86*, 837-845.