Contents lists available at ScienceDirect

Cognitive Psychology

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

Causal invariance as a tacit aspiration: Analytic knowledge of invariance functions

Jooyong Park [a], Shannon McGillivray [b], Jeffrey K. Bye [c], Patricia W. Cheng [d,*]

[a] Seoul National University, South Korea
[b] Weber State University, United States
[c] University of Minnesota, United States
[d] University of California, Los Angeles, United States

ARTICLE INFO

Keywords:
Causal induction
Causal-knowledge generalization
Causal invariance
Integration functions
Analytic knowledge

ABSTRACT

For causal knowledge to be worth learning, it must remain valid when that knowledge is applied. Because unknown background causes are potentially present, and may vary across the learning and application contexts, extricating the strength of a candidate cause requires an assumption regarding the decomposition of the observed outcome into the unobservable influences from the candidate and from background causes. Acquiring stable, useable causal knowledge is challenging when the search space of candidate causes is large, such that the reasoner's current set of candidates may fail to include a cause that generalizes well to an application context. We have hypothesized that an indispensable navigation device that shapes our causal representations toward useable knowledge involves the concept of *causal invariance* – the sameness of how a cause operates to produce an effect across contexts. Here, we tested our *causal invariance hypothesis* by making use of the distinct mathematical functions expressing causal invariance for two outcome-variable types: continuous and binary. Our hypothesis predicts that, given identical prior domain knowledge, intuitive causal judgments should vary in accord with the causal-invariance function for a reasoner's perceived outcome-variable type. The judgments are made as if the reasoner aspires to formulate causally invariant knowledge. Our experiments involved two cue-competition paradigms: blocking and overexpectation. Results show that adult humans tacitly use the appropriate causal-invariance functions for decomposition. Our analysis offers an explanation for the apparent elusiveness of the blocking effect and the adaptiveness of intuitive causal inference to the representation-dependent reality in the mind.

1. Introduction

We all have a yearning to make sense of the world, to represent causal relations between observed events in a way that finds order and meaning from overwhelming complexity. In our attempt to most effectively represent how the world works, what criteria do we use for formulating, evaluating, and revising our hypotheses? As with any form of inductive inference, there is the issue of underdetermination (e.g., Atlas, 2005): inevitably there will be multiple causal hypotheses that can explain the same observed pattern of events.

The magnitude of the issue is better revealed when we realize that reality does not come represented (Hawking & Mlodinow, 2010;

* Corresponding author at: Department of Psychology, University of California, Los Angeles, CA 90095, United States.
E-mail address: cheng@lifesci.ucla.edu (P.W. Cheng).

https://doi.org/10.1016/j.cogpsych.2021.101432
Received 21 September 2018; Received in revised form 18 August 2021; Accepted 9 September 2021

Kant, 1781/1965). Our perceptions and conceptions of reality are our representations of it, formulated within an infinite search space. Machine-vision, vision-science, and neuroscience research have brought this problem of under-determination from a quarter-millennium-old insight in philosophy into the realm of cutting-edge science, engineering, and clinical treatment (Botvinick & Cohen, 1998; Hofer, Singer, & Williams, 2005; Jayadevan, Michaux, Delp, & Pizlo, 2017; Pizlo, 2001; Mitchell & Rushton, 1971; Ramachandran & Hirstein, 1998). Consider your perception of a cube, for example, a percept that you are likely quite confident of. However, the 2-dimensional image cast by a cube on our retina is ambiguous in that it can map onto an infinite number of 3-dimensional objects, objects that need not be symmetric, and whose edges do not have to be straight lines (e.g., see Pizlo, 2001).

Yet, despite the inherent under-determination of the distal object, we are not paralyzed by indecision. We confidently perceive a cube. Narrowing down to this adaptive percept in the infinite space of possible distal objects illustrates the application of potent constraints in the form of *a priori* assumptions, in this case the default assumption that the distal object has the simplest form that is consistent with the image (i.e., the object is a "parsimonious explanation" of the image). Likewise, to construct causal beliefs in the vast search space of possible representations of reality, an adaptive solution requires cognitive constraints (Cheng & Lu, 2017).

Unlike typical current methods in artificial intelligence such as causal Bayes nets (e.g., Pearl, 2000; Spirtes, Glymour, & Scheines, 1993/2000), the human causal-induction process did not evolve in response to situations in which data input are already encoded in terms of predefined variables. The variables and concepts that serve our purposes are representations that we alone formulate—we are our own "users". This difference in the input to the causal-induction process leads to two distinct computational-level problems (Marr, 1982)—one for humans and another for current artificial-intelligence methods. Different problems require different solutions.

Recognizing the distinct problems being solved enables a reconciliation between two seemingly contradictory lines of findings. On one hand, even for the simplest causal networks, which involve three nodes, intuitive human causal reasoning violates the causal Markov assumption, a fundamental assumption in causal Bayes nets (e.g., see Meder, Hagmayer & Waldmann, 2008; Park & Sloman, 2013; Rehder, 2014, 2015; Rehder and Burnett, 2005; Rottman & Hastie, 2014, 2016). On the other hand, intuitive estimates of causal strength in three-node networks are remarkably rational (e.g., Buehner, Cheng, & Clifford, 2003; Liljeholm & Cheng, 2007; Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008; Lu, Rojas, Beckers, & Yuille, 2016; Novick & Cheng, 2004; Wu & Cheng, 1999). These estimates are consistent with the causal invariance assumption, also termed the independent-causal-influence assumption.

In the present paper, we consider the *causal invariance hypothesis*, which posits that the tacit aspiration to acquire causal knowledge that holds when it is applied is a key component of a rational solution to the natural causal-induction problem (Cheng, Liljeholm, & Sandhofer, 2013, 2017, under review; Cheng & Lu, 2017). We report two experiments testing some predictions of our hypothesis in two well-known inductive learning paradigms (Kamin, 1968; Kremer, 1978; Rescorla & Wagner, 1972).

1.1. Overview of framework

Here, we give a brief introduction to our hypothesis. In subsequent sections, we further clarify the nature of humans' causal-induction problem, and explain why causal invariance as an aspiration is essential to solving this problem.

Whether specific factors (as represented) are independent causes of an outcome or interact to form a complex cause can only be learned empirically, from one's own or someone else's observations. But, for the acquired knowledge to be useful, such empirical learning is, in our view, dependent on *analytic knowledge* of *causal-invariance integration functions*. Such knowledge, which concerns what happens when multiple causes of an outcome are present, states what the expected outcome *would be if* the causal mechanisms operate the same way in combination as when they operate alone. It does not state that the causal mechanisms *are* invariant; nor does it state what the outcome *must be* or *will be*.

When the observed outcome deviates notably from the outcome expected under causal invariance, the deviation provides a signal to the reasoner indicating a potential need to revise current empirical knowledge *toward* greater causal invariance. Such revision may include the formulation of candidate causes that are not in the current representations.

For example, reasoners may have a hypothesis that consuming fiber helps prevent metabolic syndrome. But, if they find that eating whole fruits helps lower their blood-sugar level and thus prevent metabolic syndrome, while drinking fresh fruit juices with the equivalent amount of pulp fiber does not lower (but may in fact raise) their blood-sugar level, then they would see that the preventive strength of fiber is not invariant across contexts. The deviation from invariance would be a cue to revise the original hypothesis. Whereas reasoners' current set of candidate causes may all be formulated in terms of dietary substances and their interactions, the deviation may motivate looking beyond that set of possibilities, to consider the topology of fiber as a means to re-represent the cause in a more invariant way. Specifically, fiber—in the form of intact cell walls enclosing the fructose in the cell—sequesters fructose "away from the absorptive surface of the small intestine," blocking it from being rapidly released into the digestive system (Ludwig, 2013, p. 34). This revised knowledge is more invariant in that it predicts the presence or absence of metabolic syndrome more accurately across a broader set of contexts: both fiber in whole fruits and fiber in fresh fruit juice. Furthermore, the revised knowledge generates novel predictions (e.g., eating guacamole would protect against metabolic syndrome but eating apple sauce would exacerbate it, because the crushing of cell walls would not affect fructose absorption in a low-fructose fruit like avocado but would affect it in a high-fructose fruit like apple).

Now, what is believed to be a more invariant cause may subsequently be found to be non-invariant in a new context, in which case reasoners may further revise their causal knowledge, depending on their resources. The signaling of a deviation from causal invariance

reiterates.[1] Notably, current artificial-intelligence methods, to our knowledge, do not seek to revise the user-supplied variables.

Although causal invariance is an important topic in the philosophy literature (e.g., Woodward, 2003, 2006, 2010, 2021), it has been relatively neglected in the cognitive science literature. For example, cognitive scientists studying *integration functions* or *functional forms*—mathematical functions that specify how causes combine their influences on an outcome of interest—have not considered a role for causal invariance as an aspiration or a cognitive constraint (e.g., Beckers, De Houwer, Pineño, & Miller, 2005; Beckers, Miller, De Houwer, & Urushihara, 2006; Gopnik, Griffiths, & Lucas, 2015; Griffiths & Tenenbaum, 2005; Lovibond, Been, Mitchel, Bouton, & Frohardt, 2003; Lucas, Bridgers, Griffiths, & Gopnik, 2014; Lucas & Griffiths, 2010; Melchers, Lachnit, & Shanks, 2004; Shanks & Darby, 1998; Tenenbaum & Griffiths, 2001). In these accounts, if causal invariance is considered special (e.g., the additivity function as discussed in Beckers et al., 2006), it is only as a default that is simply abandoned when it is disconfirmed by observations. Even when specific integration functions are regarded as more appropriate than other functions for certain outcome-variable types (Lu et al., 2016), no connection is made to the concept of causal invariance, which in our view explains *why* those functions are appropriate and *what* is in common across the "appropriate" functions. (Appropriateness might be due to reasoners' prior experience of how causes in a domain combine, for example.)

The neglect of causal invariance may in part stem from the belief that the aspiration to develop theories that are invariant across contexts is a hallmark of scientific reasoning (Sloman, 2005; Woodward, 2000; 2003; 2010) but not intuitive everyday causal learning. Curiously in that regard, intuitive evaluations of causation involving a binary outcome variable (the most studied outcome type in the human causal-learning literature) do assume causal invariance as a default (e.g., Buehner et al., 2003; Liljeholm & Cheng, 2007; Lu et al., 2008, 2016). Do intuitive reasoners (tacitly) aspire toward invariant causal explanations across contexts? In Experiments 1 and 2, we tested some predictions that follow if they do. Is it coincidental that causal invariance plays a role in both scientific and intuitive reasoning? Our analysis below suggests not.

1.2. The natural causal-induction problem

What is required to solve a causal-induction problem in which the search for causal knowledge occurs in the infinite space of possible representations? Given that enumerating all possible representations and evaluation them is not an option, a fruitful search for causal knowledge in that space would require a signal to tell the reasoner when to look outside their current—necessarily limited—candidate set, to formulate or reformulate concepts and variables. The signal itself is a representation, albeit that of an imagined possible world.

We assume the reasoner's goal is to acquire causal knowledge that is *useable*, in the sense that knowledge induced from prior experiences holds true when subsequently applied. However, every new situation differs in some (potentially unknown) way from situations in the past. Thus, when a reasoner applies causal knowledge, they inherently assume causal invariance, namely, that while background causes may vary, the influence of the cause in question is stable—that the causation aspect of nature does not change. As the philosopher David Hume (1748/1975, p. 38) notes, "if there is any suspicion that the course of nature may change, … all experience becomes useless." In other words, the goal of useable causal knowledge implies a reasoner-imposed constraint of *causal invariance*: the unchanging operation of a causal mechanism from a learning context to an application context. We use "*contexts*" to denote situations in which background causes of an outcome may occur; in different contexts, these causes may occur with different probabilities. By *background causes*, we mean (observed and unobserved, known and unknown) causes other than the candidate causes in question.

To see the role of causal invariance as an indispensable navigation device within the infinite search space of causal representations, consider three environmental constraints. First, causal relations are inherently unobservable (Cheng, 1997; Gopnik, Glymour, Sobel, Schulz, Kushnir, & Danks, 2004; Griffiths & Tenenbaum, 2009; Hume, 1739/1987; Ichien & Cheng, in press; Pearl, 2000; Spirtes et al., 2000). By causal relations being "unobservable", we mean that causal relations are not in the input available to a downstream construct that we label the causal-induction process. Given that our sensory input does not contain causal relations, but we "know" causal relations, there must be a downstream process that does the work of *inferring* causal relations, the output, from its noncausal input. What *is* "observable", namely, the input to that process, are the states of candidate causes (as represented) and the aggregate outcome due to *all* contributing causes (observed and unobserved causes). The orbit of the Moon we observe, for example, is the orbit due to the total gravitational pull from all celestial bodies. No telescope will delineate separate orbits due to the Sun, the Earth, etc., no matter how powerful! These individual causal powers must be inferred through the causal-induction process.

Second, the environment always has the potential to contain unknown background causes of an outcome (hence the mantra "association need not imply causation"). Assessing the embedded potential contribution of a candidate cause therefore requires extricating it from the influence of the background causes. This crucial step during learning requires a *decomposition* assumption specifying how the observed outcome can be analyzed into influences from the target cause and those from the contextual background causes. Nature does not tell us how to do the extrication: the task is reserved for us the reasoners.

Third, it is always possible for background causes of an outcome to differ across contexts. Because events in the world inevitably change, the injunction in statistical textbooks for scientific inference to only interpolate (i.e., to restrict inference to the population sampled) is unrealizable in practice. The universe we inhabit has never repeated itself. By the time we apply causal knowledge from a medical study, the population from which the study drew its sample has aged; climate change may have brought pervasive and

[1] Our present paper addresses *when* causal knowledge needs revision but not *how* it is revised. How it is revised would depend on the domain and the reasoners' knowledge and imagination.

unanticipated health consequences; societal forces such as new technology, immigration, and the interconnectedness between human populations might have modified the younger population's habits, culture, and genetics. For everyday causal inference, which does not involve random sampling of a population, interpolation is not even a theoretical possibility—there is no representative, randomly sampled population to which to circumscribe interpolative inference. Extrapolation is our only choice.

Predicting the expected outcome in a new context by applying previously induced causal knowledge requires a *composition* assumption specifying how to combine the influences of the candidate and (new) background causes. Composition is the inverse of decomposition. (We refer to both functions characterizing decomposition and composition as integration functions.)

Given these objective constraints, how should causal relations be extricated from their incidental context? Does the subjective goal of formulating useable causal representations impose a constraint on the extrication? Our paper addresses this issue at the level of *what* is computed rather than *how* it is computed (Marr, 1982).[2]

2. An assumption in the construction of generalizable causal knowledge

2.1. Causal invariance as aspiration

Our proposal is that the concept of causal invariance shapes our representations by serving as a default decomposition function, a default composition function, and a knowledge-revision criterion, with the three roles all driven by the (tacit) goal to acquire causal knowledge that *generalizes across contexts*, in particular across the learning and application contexts (e.g., Cheng et al., 2013, 2017; Cheng & Lu, 2017). Because that goal is likely never reached across all contexts (consider theory revisions in the history of science, Kuhn, 1962/2012), we refer to it as an aspiration.

To explain the three intimately related roles in this subjective constraint, we make two distinctions: 1) a part-whole distinction, between a "*whole*" cause (elemental or complex) and an *interactive component* (part) within a whole cause, and 2) a distinction between *analytic* and *empirical* knowledge (cf. Hume's, 1739, fork of knowledge: "relations of ideas" and "matters of fact"). Stating the goal of causal learning in terms of these distinctions, *the goal is to formulate whole causes—potentially consisting of interacting components—using empirical information from our observations, such that the whole causes would conform to how our analytic knowledge of causal invariance informs us they should behave in application contexts that may have different background causes; that is, such that whole causes are ideally teased apart from (do not interact with) other causes* (e.g., those in the background). We explain each distinction in turn below.

Consider examples of the part-whole distinction. To "force quit" a frozen computer program, a user must simultaneously press multiple keys, such as Ctrl + Alt + Delete. Each of these key presses is a necessary component of a "whole" cause of the outcome. Playing a guitar chord requires a conjunction of finger placements and strumming. To have "healthy forest growth," there must be adequate nitrates in the forest soil, rain, drainage, amounts of carbon dioxide and of oxygen, elevation, sunshine, etc. The conjunction of these individually insufficient but necessary factors is a complex "whole cause" of growth[3], alongside other whole causes of forest growth, such as logging by the lumber industry. Each whole cause (whether elemental or complex, generative or preventive) by itself has an influence on growth (cf. Mackie's, 1974, "*INUS*" condition: each component factor is an insufficient but necessary part of a condition which is itself an unnecessary but sufficient cause of the outcome).

Our second distinction concerns the basis for belief. Whereas analytic knowledge is justified by *reason*—that is, by what logically follows from a set of premises given the meaning of the concepts in question—empirical knowledge is justified by *experience*, by encounters with the world.[4] Let us use an example from Luria (1931) to illustrate the distinction. A villager in Uzbekistan was asked, "In the far North, where there is snow, all bears are white. Novaya Zemlya is in the far North, and there is always snow there. What color are the bears there?" One of the Uzbek villagers replied, "Your words can be answered only by someone who was there." To this villager, the answer is to be settled by testimony based on experience, and by that only. They might have wondered why Luria asked such an odd question. By contrast, an analytic answer would be a conditional: *If* Luria's premises are true, *then* the bears in Novaya Zemlya are white.

In the case of integration functions, empirical knowledge of an integration function would state (after actual encounters) that causal influences *a*, *b*, and *c* combine as *described* by function *f* to produce outcome *x* on the basis of observations. By contrast, analytic knowledge of a causal invariance integration function would be a conditional that states the expected outcome under its premise: *if* causes are invariant across contexts, *then* the outcome would be as predicted according to < the appropriate causal invariance function for that outcome variable type > . The premise of the conditional is the *sameness* of the operation of *a causal mechanism across contexts* given the meaning of "sameness", "causal mechanism", and "context". Such analytic knowledge is necessary because in order to discern whether an observed outcome indicates a causal interaction, one must first specify the outcome that *means* "no interaction", then compare that criterion to the observed outcome.

To consolidate our part-whole and empirical-analytic distinctions, let us illustrate both distinctions within Newton's theory (1687/

[2] In other words, our paper does not address how the inferences are implemented, by propositional reasoning, mathematical reasoning, or some other process.

[3] The conjunctive cause consisting of the combination may be regarded as a place-holder for more in-depth revisions. A botanist, for example, may further explain why the conjunction of sunshine, water, and nutrients in the soil produces plant growth.

[4] Analytic knowledge can be either qualitative or quantitative. Conversely, quantitative knowledge can be either analytic or empirical. The formula in Newton's (1687/1713) inverse-square law governing celestial and terrestrial motion, for example, is quantitative, but nonetheless empirical because it is justified by observations.

1713) of celestial motion. Newton's law of universal gravitation characterizes the "whole cause"—the vectors representing the equal and opposite gravitational pull at any moment between any two bodies—as an interaction among its component factors (as a direct function of the masses of the two bodies and an inverse function of the square of the distance between them). This inverse-square law is empirically induced. The causal-invariance integration function is vector addition. In our view, the gravitational forces obeying the empirical function nicely hold between any two bodies regardless of other bodies in the context *because* the variables in that law were *formulated so that* any individual "whole cause" would be invariant across all possible configurations of other bodies in the context, with vector addition as the analytic expression of that invariance. More generally, the reasoner (to the extent resources are available) seeks to represent causes that are invariant across contexts.

To see how the aspiration of causal invariance shapes causal representations toward useable causal knowledge, we can make use of the symmetry of the learning and application contexts. It seems that the minimal requirement for a causal relation to be useable would be that it holds in the original context in which it was inferred. As explained earlier (Section 1.2), when applying causal knowledge in a new context to predict the outcome, the default composition function for the target cause and the contextual causes is inherently causal invariance. But, to satisfy that minimal requirement, note that during the original learning, causal invariance must also be the default integration function that *decomposes* the observed outcome into the contributions from the candidate cause and the background. Take a causal relation that has been found to generalize to an application context (using causal invariance as the composition function to correctly predict the outcome). Now, treat this as a "learning" context, and consider "applying" this successfully generalized relation back in the original context. Which context is encountered first is incidental and should not affect the content of the inferred causal relation that, under our supposition, holds across the two contexts; flipping the "learning" and "application" context labels should make no difference. But, composition is the inverse of decomposition. Thus, if causal invariance had not been the decomposition function during the original learning, the hypothetical flipped test—which applies the target causal relation using causal invariance now as the *composition* function—would not correctly "predict" the actual data in the original learning context. (See Ichien and Cheng, in press, for data-set examples.)

In other words, for any causal mechanism that indeed remains unchanged across two contexts, the desired inferred causal relation that does generalize across the contexts could not have resulted from the actual data in the learning context *unless* causal invariance is the decomposition function during learning.

Intimately tied to causal invariance's role as a default integration function is its role as a revision criterion: deviation from causal invariance serves as a signal to revise causal knowledge. What if the need for revision were signaled instead by deviation from a *causal-interaction* (i.e., non-causal-invariance) criterion? If candidate c of effect e is expected to interact with some component of the background causes, then there would be 1) a deviation from this expectation—signaling a need to revise causal knowledge—when the actual influence of c in fact generalizes across contexts (resulting in a false alarm), and 2) no deviation from expectation—hence no signal to revise— when c interacts with background causes as expected (resulting in a miss). If c interacts with background causes, the causal relation as represented can actually benefit from revision, because it presumably would not generalize to other contexts, as not all contexts contain the same interacting component. In other words, adopting causal interaction as a revision criterion would result in faulty signals. Because confirmation and disconfirmation depend on the signal, such systematic errors would be grievous in an immense search space.

In summary, to enable generalization across contexts where unknown causes may be present, the aim of causal-knowledge construction is to formulate whole causes that are teased apart from other causes. Given the infinite space of possible representations, whole causes (e.g., Newtonian gravitational force) are not independent of other whole causes by happenstance: they are representations that were constructed expressly to attain causal invariance. Our analysis explains why causal invariance plays a role in both scientific and intuitive causal learning. In our framework, scientists aspire to causal invariance because they share with lay folks a deep-rooted solution to the natural causal-induction problem in the search space of reality where variables await our representation.

2.2. A contrast with empirical knowledge of integration functions

What is the scope with which empirically acquired integration functions generalize? For example, do such functions generalize within a domain (celestial motion but not forest growth)? Do they generalize within a variable type? We start with some examples, then consider the general issue.

Within the domain of celestial motion, we see that whereas an empirical interaction function (the inverse-square law in Newton's, 1687/1713, theory) combines the interactive influences from the component factors, a causal-invariance function (vector addition) combines the independent influences from multiple whole causes (e.g., the force vectors representing the gravitational pull on the Moon from the Sun, the Earth, and other bodies). Notice that the four variables in Newton's law (the gravitational pull between two bodies, the masses of the two bodies, and the distance between them) may all be regarded as being of the continuous type. This example illustrates that, for variables of the same type in the exact same domain, an interaction among some variables co-exists with invariance among other variables. We see the same co-existence for the continuous variables in the forest growth domain. As our examples illustrate, empirically learned interaction functions generalize neither across variables of the same type nor across variables within a domain.

Such failure to generalize empirical integration functions is to be expected. Under our causal invariance view, if our empirical knowledge is successfully formulated (i.e., invariant and useable), the interactions among component factors should not generalize to variables at the whole-cause level! The interaction among factors such as sunlight and soil nitrogen in the conjunctive cause that produces healthy forest growth should not generalize to the relation between that conjunctive whole cause and logging by the lumber industry. (Because conjunctive causes are common, there would be more interactive-component causes than whole causes within the

same domain.) That is, empirical integration functions are specific to the *content* of the represented variables (e.g., sunlight, soil nitrogen, forest growth; mass, distance, gravitational force) rather than to their *form* (continuous scalar, vector, etc.). Whether and how they generalize depends on the similarity between the underlying causal mechanisms in question (e.g., Lucas & Griffiths, 2010, Expt. 5; Miller & Schachtman, 1985; Wheeler, Miller, and Beckers, 2008, Expt. 3) as well as on situational variables (e.g., the demand characteristics in an experiment; Wheeler et al., 2008, Expts. 1 & 2). Notably, although an empirical interaction function may serve as a default hypothesis for variables that are perceived to share the same causal mechanism (e.g., novel cues in the same domain as learned conjunctive causes; e.g., Melchers et al., 2004), if that default in fact fails, that interaction function is simply abandoned, and is not retained as a goal for knowledge representation.

In contrast, analytic knowledge of causal-invariance functions (e.g., vector addition, the "noisy-OR" function in Eq. 1 below, wave addition) is formal, that is, domain-, content-, and context-independent: invariance functions are specific to outcome variable types (e.g., vectors, binary variables, waves), with their validity resting on logical consistency given the meaning of "sameness" as applied to the outcome-variable type of the causal mechanism in question. The *validity* of the if-then conditional,

> *if* < cause x > influences outcome z *the same way* regardless of the state of < cause y>, *then* it follows that their combined influence would be as specified by < the causal-invariance function for outcome type z given the states of causes x and y>,

does not depend on the *truth* of the consequent: a mismatch between an observed outcome and the consequent of the conditional only implies (by the contrapositive) that the antecedent is false (which may trigger revising the hypothesis relating x, y, and z). Thus, vector addition would always represent causal invariance for vectors, and would *remain* a goal for knowledge revision whether or not a specific predicted invariance actually holds.

2.3. Analytic knowledge of causal invariance functions: For continuous and binary outcomes

Given that analytic knowledge of causal-invariance functions enables computing the expected outcome assuming *superposition*—the combination of invariant causal influences—reasoners must have such knowledge, at least in an implicit form, for at least some outcome-variable types, to construct new useable causal knowledge involving those outcome-variable types. Notably, the superposition of causal capacities is manifested as distinct mathematical functions for different outcome variable types. For example, superposition is expressed for a binary outcome (e.g., a bone being fractured or not, a patient being alive or dead) by noisy-OR and noisy-AND-NOT functions; for a continuous outcome (e.g., food in varying quantities, the amount of light from multiple lamps falling on an area per unit time) by scalar addition; and for vectors (e.g., gravitational or electric force vectors) by vector addition.

For scalar continuous (and approximately continuous) outcome variables, in the range where an outcome Z is on an interval scale and not at a ceiling or floor level, changes in the strength of a cause produce corresponding changes in the intensity of Z. Thus, *if* the magnitude of change in Z produced by each of two generative causes is unaffected by the state of the other cause (i.e., their causal powers combine invariantly), *then* the magnitude of change in Z when these two causes both occur would be the arithmetic sum of the magnitudes of change in Z produced by each of the respective causes alone; that is, causal invariance manifests itself as *additivity* for scalar continuous outcomes. Consider the physical amount of light cast in an area (rather than the area's perceived brightness):

If two lamps each casts 1000 lm of light when it is on by itself,
and each lamp *does not change* the amount of light energy it casts *depending on* the state of the other lamp,
then together the area would receive 2000 lm of light when both lamps are on.

As should be clear, the validity of this if-then conditional is justified by reason (cf. Luria's, 1930, question about white bears). In particular, the validity of the conditional is distinct from the truth of the consequent: an observation that the consequent is false (e.g., the area receives a total of 1500 lm rather than 2000 when both lamps are on) invalidates not the conditional itself but rather the assumption that the lamps cast light invariantly (i.e., denying the consequent of a conditional entails that the antecedent is false, per *modus tollens*). Say, the lamps are energy-conserving and automatically dim when they detect an ambient light beyond some brightness threshold. The combined outcome would be less than 2000 lm in that case, deviating from causal invariance, but the conditional implementing additivity as the causal-invariance function would still hold.

In contrast to continuous variables, binary variables have no graded intensity levels. For such variables, magnitude of causal capacity is expressed by probabilities. Generative causal capacity, for example, can be represented as the probability of a cause producing the effect in an entity (Cartwright, 1989): The *capacity* of cause C to produce effect E in an entity is the probability of E occurring in the entity in the presence of C when all other causes of E are absent (potentially counterfactually). Causal capacities, being unobservable, must be estimated by inference under a set of assumptions from observed frequency data for the occurrences of the cause and effect variables. Note that the relative frequency of occurrence of an outcome in a set of entities is a proportion involving cardinal numbers (e.g., 3 out of 10 patients treated have a fractured bone), which are on a ratio scale. Units on this scale have a physical meaning and are not arbitrary.

For binary outcome variables, in accord with the product definition of independence in probability theory (e.g., Feller, 1957/1968; Jaynes, 2003), the *sameness* of the causal capacities of generative causes of an outcome manifests itself as a noisy-OR integration function (e.g., Cheng, 1997; Glymour, 2001; Good, 1961; Pearl, 1988; Sheps, 1958; Yuille & Lu, 2008). For generative causes A and B, assuming they are the only causes of E, and the capacity of each cause to produce E is unaffected by the other cause (i.e., they are invariant), the *noisy-OR* function states that the probability of E occurring is:

$$P(e = 1|a, b; p_A, p_B) = p_A a + p_B b - p_A p_B ab,$$

where a, b, $e \in \{0, 1\}$ denote the absence and presence of A, B, and E, and p_A and p_B denote the respective capacities of A and B to produce E. Note that the sub-additivity of causal invariance for binary outcomes is **not** a ceiling effect: the noisy-OR function deviates from arithmetic additivity whenever both A and B produce the outcome with a nonzero probability (see the negative product term in Eq. 1). We return to this point in Experiment 2.

If our analysis is correct, people should intuitively use the appropriate causal-invariance functions for at least some outcome variable types. Our experiments tested whether untutored reasoners use the appropriate causal invariance functions for continuous and binary outcome variables.

2.4. An alternative view: The purely empirical integration-rule-learning hypothesis

One might argue that the reported use of the additive function for continuous outcomes and the noisy-OR function for binary outcomes are empirically justified, and are *not* interpreted analytically in terms of the sameness of causal influences across contexts. Instead, reasoners might have merely learned from their experience that—as a matter of empirical fact—most causal variables of continuous outcomes combine additively, whereas most causal variables of binary outcomes combine in a noisy-logical manner. Note that because the additive and noisy-OR integration functions are in fact causal invariance functions for their respective outcome variable types, the *empirical rule-learning hypothesis* implies that most of our individually hypothesized and culturally transmitted causal relations have elemental causes (with no interacting components) that do generalize across contexts.

This hypothesis is both conceptually and empirically problematic. Fundamentally, the empirical rule-learning hypothesis is inconsistent with a basic tenet of cognitive science: namely, that our conception of the world is our representation of reality rather than reality itself. There are many possible representations of reality, for something as mundane as the color of an object (cf. the artist James Turrell's Skyspaces) or as profound as the nature of time and space. The empirical rule-learning hypothesis overlooks the crucial fact that which integration function characterizes how variables are observed to combine depends on the chosen representation of those variables.

Consider the representations of planetary motion. Isaac Newton's choice of variables enabled his universal law of gravitation to yield gravitational forces that superpose on each other. Not all choices of variables enable that desirable property. If Newton had instead chosen other variables (e.g., "conatus", a body's inherent tendency to move instead of a body's mass, as proposed by Hobbes and Leibniz, or a circularly moving body's inherent "center-fleeing" (centrifugal) tendency, as proposed by Descartes), superposition would not have resulted. Needless to say, there are any number of other possible representations. It took a Newton to have formulated the variables in his universal law and framework. (A simple concrete example in our general discussion further illustrates how choice of representation affects whether invariance results.) The hypothesis that causal-invariance functions are empirically learned implies, anomalously, that regardless of the choice of representation, the chosen variables would typically give invariant causal knowledge.

Moreover, the empirical rule-learning hypothesis implies that if most chosen variables fail to generalize across contexts (i.e., interact with background causes), then interaction-integration functions would be the adopted default. In fact, many causes in our knowledge base involve complex interacting components. Consider traits such as an individual's height or weight. Each trait depends on the interaction of a large number of variables: genes, societal factors such as the wealth of one's socio-economic group during childhood, the influence of the agricultural and food-processing industries, the individual's food preferences and habits, etc. Likewise, whether someone has a car accident is likely the result of the interaction of numerous factors, each one of which would not have produced the outcome on its own. In our gravitational force and forest growth examples, for each whole cause there are three or more component causal factors that combine interactively. Thus, it seems that the burden of proof is on the empirical rule-learning hypothesis to demonstrate that most causes we have inferred are represented elementally.

Even if it were the case that causal knowledge inherited from our current culture tends to consist of elemental invariant causes, why would human representations have begun in terms of conveniently generalizable variables right from the start? In fact, archaeology and history inform us that human causal knowledge has evolved enormously (e.g., see Diamond, 1997; Kuhn, 1962/2012; Needham, 1956, 1962, 1965, 1985a–c, 2000). Given that the intuitive human causal-learning process is unlikely to have undergone fundamental change within the same time frame, that process must be capable of freedom from the rut of failures, capable of formulating (relatively) invariant causal relations despite having proposed and tested numerous hypotheses that fail to generalize. In sum, from multiple perspectives, the empirical rule-learning hypothesis is untenable.

Turning the empirical rule-learning view on its head, our account holds that the (relatively) invariant causes that are useful to our understanding of the world are formulated with invariance as an aspiration rather than by happenstance. Rather than the world presenting itself to us already represented, the human causal-inference process develops the (potentially complex) invariant "whole" causes drawing on analytic knowledge of causal invariance functions, which differ mathematically depending on the form rather than the content of the outcome variable, as a tool to navigate the vast search space of representations (Carroll & Cheng, 2010; Cheng & Lu, 2017).

3. Empirical tests of the use of causal-invariance functions

3.1. Overview of experiments

We present two experiments testing whether people use the appropriate causal invariance functions as a criterion for hypothesis revision. If reasoners aspire toward causal invariance, and they have analytic knowledge of the causal invariance functions for continuous and binary outcomes, they would use the respective causal-invariance function in accord with an outcome that is

represented as continuous or binary.

Experiment 1 tested whether reasoners' causal judgments in a blocking paradigm (Kamin, 1968) are consistent with their spontaneous use of the respective causal-invariance functions for continuous and binary outcomes, depending on their interpretation of the outcome variable type, which our experiment manipulated. To our knowledge, no previous study has manipulated the interpretation of the outcome variable type while keeping all else—including prior domain knowledge and the actual outcomes—constant. Experiment 2 tested the appropriate use of the causal-invariance function for a binary outcome (Eq. 1) in the overexpectation paradigm (Kremer, 1978; Rescorla & Wagner, 1972), evaluating predictions based on noncausal integration functions against those based on Eq. 1. Experiment 2 additionally clarified the causal nature of the application of Eq. 1. In particular, it tested the role of the "no confounding" prerequisite for causal induction in the application of that equation.

Our experiments test qualitative predictions of the causal-invariance hypothesis, and do not require that reasoners have explicit knowledge of the relevant invariance functions. Although knowledge of the noisy-logical functions can be made explicit (as we did in Eq. 1), people's use of these functions is likely to draw on knowledge that is implicit. We return to this issue in the general discussion. Both our experiments obtained informed consent from participants.

3.2. Experiment 1

Experiment 1 utilized a blocking paradigm (Kamin, 1968) adapted for human participants from the non-human associative learning literature. We first explain the typical blocking paradigm and subsequently our adaptation. In the usual paradigm, participants are shown multiple instances in which stimulus cue A is paired with the target outcome. Following conventional notation, we label such trials as A+ trials ("+" indicates the occurrence of the target outcome). Subsequently, cue A is repeatedly shown co-occurring with cue X, and this combination of the two cues is paired with the outcome (AX+ trials). Thus, cue X, like cue A, is always paired with the presence of the outcome. Cue X is the cue of theoretical interest. The control cues for X are two other cues, B and Y, that only occur in combination, and are always paired with the outcome (BY+). Note that blocking training does not convey any information on the integration function because only cue A is ever presented alone.

During a subsequent test phase, participants rate how likely each cue alone is to cause the target outcome. Because cue X only differs from the control cues (B and Y) in that X was paired with a cue that was previously associated with the outcome on its own (cue A), any reduction in the causal rating for cue X relative to those for the control cues can only be due to the acquired knowledge regarding cue A. We operationally define the *blocking effect* to refer to this reduction due to knowledge acquired during the prior A+ trials. In an associative framework, the effect may be viewed as the "blocking" by the prior A+ trials of the accumulation of associative strength between X and + despite their constant pairing during the AX+ trials. The amount of blocking is the extent of this reduction in the causal rating for X. In our experiment, the target outcome was an allergic reaction, and cues A, B, X, and Y were assigned to four potential food allergens, with the assignment counterbalanced across participants.

To disambiguate the variable type of the target outcome, our Experiment 1 used a novel variant of the just-described paradigm. Prior to blocking training, our experiment introduced a short *pre-training phase* in which one and only one stimulus cue appeared: ragweed pollen. Ragweed pollen appeared in this phase only, but the target outcome was the same allergic reaction as in the subsequent blocking training phase. The level of exposure to the ragweed pollen was indicated by a pollen index: a continuous variable. We manipulated the interpretation of the outcome-variable type by randomly assigning participants to one of two between-subjects *outcome-variable-type conditions*: increasing levels of the pollen index (kept identical across conditions) were paired with *either* a) increasing levels of the allergic reaction, as indicated by an increasing number of red dots on a patient's face (the *Continuous condition*), or b) with a single level of the allergic reaction whenever a threshold was exceeded (the *Binary condition*). The single level in the Binary condition was identical to the maximum level in the Continuous condition (the same number of red dots appeared on the patient's face). All participants therefore experienced the exact same range of the allergic reaction. The pairing between levels of ragweed pollen with levels of the allergic reaction in this pre-training phase was the only difference between conditions. The purpose of the pairing was to disambiguate the nature of the outcome variable, as continuous or binary, notably without conveying any information on how multiple causes combine their influences (ragweed pollen was the only pre-training-phase cue).

Blocking training, the main part of the experiment, was identical across conditions. Potential allergens during blocking training were all food cues. All participants experienced identical A+ and AX+ trials with the outcome occurring at the same level — the maximal level participants experienced for both conditions during the pre-training phase. Now, by our causal-invariance hypothesis, given reasoners' *a priori* concepts of continuous and binary variables, participants in the Continuous group would infer that, according to the additive invariance function for a continuous outcome, *if* Cue X *were* causal, then it *would have* led to a stronger allergic reaction on the AX+ trials than on the A+ trials. Cue X must therefore be noncausal (i.e., X is considered "blocked" from receiving associative weight in the associative literature, although we would say the *inference* of noncausality actually "comes through"). In contrast, participants in the Binary group would infer that, according to the noisy-OR invariance function, regardless of whether Cue X is causal, AX+ trials would have the same level of reaction as A+ trials. Thus, based on the respective causal-invariance functions for the two outcome-variable types, our hypothesis predicts that Cue X should be judged noncausal in the Continuous group, but causal (with uncertainty) in the Binary group.

Note that in a blocking paradigm involving a continuous outcome, reasoners who judge Cue X as noncausal may be adhering to either causal invariance or merely to additivity (e.g., see Lu et al., 2016), without interpreting additivity as causal invariance. The judgment is consistent with either hypothesis. However, if reasoners also use the additivity function when the paradigm involves a binary outcome, then we can infer that they are adopting additivity *per se* (as in Rescorla and Wagner's, 1972, model). Experiment 1 serves to disambiguate whether reasoners adhere to causal invariance or to a particular mathematical function, by testing whether they

use different causal-invariance functions as appropriate for the respective outcome-variable types.

3.2.1. Ruling out two alternative explanations: Going beyond previous findings

To provide support for our analytically-based account, our experiment must rule out explanations due to reasoners' prior knowledge, in other words, empirically-based explanations. In previous research, participants using a noisy-logical function (e.g., as shown in Buehner et al., 2003; Liljeholm & Cheng, 2007) might have interpreted the function simply as an empirical description of the way causes in a specific domain combine their influences on an outcome (due to their prior knowledge of the domain evoked by the cover story), rather than as an expression of causal invariance. Therefore, while our experiment manipulated participants' interpretation of the outcome-variable type, it kept all domain knowledge and prior experience with integration functions constant across conditions.

Another potential empirical explanation to be ruled out is the *inferential reasoning account* of blocking (Beckers, et al., 2005; Beckers et al., 2006; Blanco, Baeyens, & Beckers, 2014; De Houwer et al., 2002; De Houwer et al., 2005; Lovibond, 2003; Lovibond et al., 2003; Mitchell, De Houwer, & Lovibond, 2009; Mitchell & Lovibond, 2002; Wheeler et al., 2008).[5] According to this account, "in order to logically infer that X is not a cause of the outcome, people not only have to assume that effective causes have additive effects, but they also have to be able to *empirically verify* [italics ours] that adding X does not increase the outcome over that produced by A. … Such verification is … prevented if the outcome is … at the maximal strength when only A is present" (Beckers et al., 2005, p. 239).

Note that this explanation of the blocking effect assumes an outcome that can occur at varying positive levels, such as a continuous outcome, so that the outcome occurring at a higher level than that produced by A is a possibility and potentially verifiable.

Similarly, Beckers et al. (2006, p. 93) write, "According to a causal reasoning analysis, blocking should be constrained not only by whether people assume additivity but also by whether people are actually able to empirically verify on AX trials that X does not add to the outcome of A …. Assume that food A results in an allergic reaction with an intensity that corresponds to the maximal intensity …. If later the combination of food A and food X results in a similar allergic reaction, the causal nonefficacy of X cannot be validly inferred, … [and] people should be unsure about the causal status of X." Note that the passage refers to the intensity of the outcome, which also indicates that the account concerns a continuous outcome variable.

Thus, according to the empirical-verification explanation, for the blocking effect that involves a continuous outcome to occur, it would not be sufficient for the reasoner to merely allow for the *possibility* that the outcome could occur at a higher level than when only A is present. That possibility needs to be "verified" by actually experiencing or having experienced a higher outcome.

Our Experiment 1 prevented the empirical verification that Beckers et al. (2005) argue is necessary by holding the outcome for both the Continuous and Binary groups at the maximally observed strength when only A is present. The empirical-inferential account therefore predicts no difference between groups: neither group should show a blocking effect because neither group has information to empirically verify that the outcome could have occurred at a higher level than when only A is present. Crucially, this account of blocking contrasts with ours not in whether inference is involved, but in whether use of analytic knowledge associated with binary and continuous outcome-variable types *is sufficient to produce a difference in the blocking effect, controlling for empirical learning and verification*.

3.2.2. Method

3.2.2.1. Participants. Forty-eight undergraduate students (average age = 19.8) at the University of California, Los Angeles participated in the experiment for course credit. The participants were randomly assigned to one of two conditions: 24 each in the continuous and Binary conditions. We recruited 24 because our experimental design has 8 counterbalancing subgroups in each condition, and we chose *a priori* to have 3 participants randomly assigned to each subgroup based on the variability of causal judgments observed in previous studies (e.g., Liljeholm & Cheng, 2007). Data from all 24 participants were included in our analyses.

3.2.2.2. Materials and stimuli. All participants were asked to imagine they were an allergist's assistant, and after reviewing information about a fictitious patient, they had to determine what potential allergens were likely to elicit an allergic reaction in the patient. Four potential allergens were presented during blocking training: milk, peanuts, mushrooms, and strawberries. The information given about each of the four food cues indicated what (e.g., milk) and how much (one serving, for all blocking training trials) the patient had consumed, along with a picture of the cue (e.g., a glass of milk). The outcome—an allergic reaction—was visually represented by a picture of a patient's face with red dots on it (see Fig. 1 for sample stimuli). The stimuli were presented via a program on a computer.

3.2.2.3. Design and procedure. All participants completed 3 blocks of trials during the experiment: one pre-training block followed by 2 training blocks. Learning was assessed at the end of each block.

As mentioned, the Continuous and Binary conditions differed from one another only in the pre-training phase. During pre-training, participants in both groups were told that the patient had been exposed to various levels of ragweed pollen as indicated by a ragweed

[5] The papers just cited often use 'additive' *qualitatively* to mean simply that the outcome due to the combined influence of multiple cues is *greater* than the outcome due to an individual cue. Mitchell and Lovibond (2002, p. 312) make this usage explicit. However, these papers other times use "additive" in a quantitative sense, for example, as when Beckers et al. (2006, p. 98) discuss additivity as a "hard-wired feature of associative learning" (e.g., Rescorla & Wagner, 1972, which uses quantitative additivity). In the present paper, we use additivity in the quantitative sense (Eq. (3)), but we test differences in the predictions of the quantitative models (Eq. (2) vs 3) qualitatively.

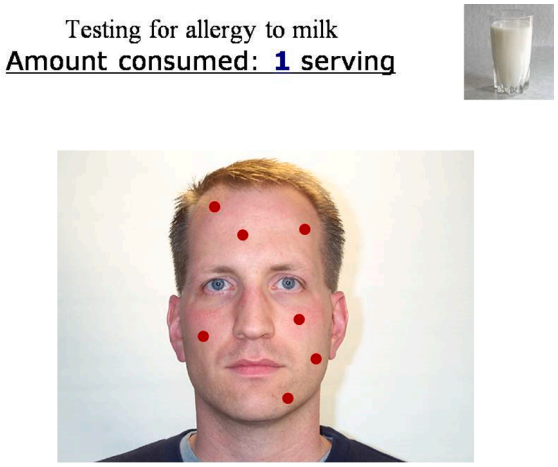

Fig. 1. Sample A+ stimulus. During presentation, the verbal potential allergen information as well as the picture of the potential allergen(s) were shown for 2 s at the top of the screen. The face then appeared below (as shown) and remained on the computer screen along with the allergen information for an additional 4 s.

pollen index (index = 0, 2, 4, 9, 17, 22, or 31; an index of 0 was presented twice, for a total of 8 trials). For the Continuous group, as the index increased, so did the number of dots on the patient's face, varying from 0 to 7 incrementally except for omitting 3. For the Binary group, the patient's face had either 0 or 7 dots depending on whether the ragweed pollen index crossed a threshold—the face displayed 0 dots when the index was 4 or below, and 7 dots when the index was above 4. Both groups were thus exposed to the same maximal outcome (7 dots on the patient's face). The pre-training trials occurred in a randomized order for each participant.

The groups completed the same blocking training, with the outcome either present at the maximal level (7 dots on the face) or absent (0 dots on the face). There were four training trial types: A+, B−, AX+, and BY−, all involving food cues. In this notation, each letter indicates a patient's consumption of a specific food assigned to that role in the blocking paradigm; "+" indicates the subsequent occurrence of an allergic reaction following the assigned food (at the maximal level); and "−" indicates no allergic reaction. During one of the two training blocks, participants were presented with 4 trials of A+ and 4 trials of AX+ . During the other training block, the control block, participants were presented with 4 trials of B− and 4 trials of BY−. As in the standard blocking paradigm, our design was *deterministic*, in that a specific stimulus cue or cue combination was either always paired with the outcome or never paired with the outcome. The trials in each block were presented in a randomized order for each participant. Cue X was the critical cue for which the empirical and analytic views make different predictions. Cue Y was the control cue. Neither X nor Y ever occurred by themselves, and their causal status therefore held the same ambiguity.

The order of the blocking training blocks was counterbalanced between participants. The pairing of the potential food allergens with stimulus cue roles A, X, B, and Y was also counterbalanced across participants. The order and pairing were independently counterbalanced, resulting in 8 counterbalancing subgroups for each condition.

Verbal instructions for the Continuous and Binary conditions were identical. After the instructions, the participants completed the pre-training block, at the end of which they were asked to assess the causal strength of ragweed pollen. Specifically, they were asked to rate how likely it would be for the patient to develop an allergic reaction when exposed to ragweed pollen with an index of 30 (near the maximal level observed) on a scale from 0 (extremely unlikely) to 100 (extremely likely). Likewise, following each of the two training blocks (either A+ and AX+ or B− and BY−), participants answered causal-strength questions regarding each potential allergen depicted during that block. For each of the food cues, participants were asked to assess how likely (on the 0 to 100 scale) the patient would develop an allergic reaction after consuming one serving of the food by itself.

Upon completion of the experiment, all participants were asked two questions that served as manipulation checks. The first checked whether participants perceived the allergy outcome as constant across the A+ and AX+ trials: it asked whether they noticed a difference in the number of dots on the patient's face when he had consumed one versus two food items (the correct answer being "no", it was always either 0 or 7 regardless of the number of food items or condition). Almost all participants in both conditions (22 of the 24 in the Continuous condition; 23 of the 24 in the Binary condition) answered correctly, suggesting that the allergy outcomes were indeed perceived as constant across the blocking trials. The second question asked participants whether they noticed any pattern between the level of the ragweed pollen index and the number of dots on the patient's face, and to explain their answer. Whereas most participants in the Continuous condition (21 of the 24) indicated noticing that the higher the level of ragweed pollen, the more red dots on the face, few participants in the Binary group (3 of the 24) gave a similar explanation. The latter group instead indicated a threshold level of exposure. The two groups' responses confirmed that our variable-type manipulation was effective: the Continuous and Binary groups perceived the two outcome variable types as we intended.

3.2.2.3.1. Predictions. All previous accounts of blocking predict no difference in the blocking effect between conditions, because 1) the cover story, blocking training, and test phases were identical across conditions, 2) our experiment supplied no information on integration functions, and 3) the outcome during blocking training always occurred at the maximum level experienced during pre-

training.

In contrast, analytic knowledge of causal invariance for different outcome-variable types predicts a bigger blocking effect in the Continuous than in the Binary condition. For the experimental variables in our Continuous condition, let w_a, w_b, w_x, and w_y respectively represent the causal capacities or weights of cues A, B, X, and Y to produce the allergic reaction, and let m represent the intensity of the reaction during blocking training (equal to the maximal level experienced during pre-training). Assuming that the cues combine their strengths additively, in accord with causal invariance for continuous outcomes, cue X should have a causal strength of 0: given that $w_a = m$ and $w_a + w_x = m$, it follows that $w_x = 0$. Because the control cue Y also has a causal strength of 0 (given that $w_b = 0$ and $w_b + w_y = 0$, it follows that $w_y = 0$), the ratings for cue X and cue Y should be highly similar. That is, cue X should be completely "blocked" by cue A.

For the Binary condition, generative causal power is the maximum-likelihood estimator of the strength (i.e., probability) of a candidate cause to produce a binary outcome (Griffiths & Tenenbaum, 2005; Tenenbaum & Griffiths, 2001), assuming a) the candidate and the background cause (a composite consisting of all other causes of the outcome) combine their (generative) strengths according to the noisy-OR function, and b) there is "no confounding" by the background cause (i.e., background causes occur just as often in the presence of a candidate cause C as in its absence, Cheng, 1997). Let c, $e \in \{0, 1\}$ denote the absence and presence of candidate C and effect E, respectively. Under these conditions, for situations in which C is potentially generative (i.e., effect E occurs at least as often in the presence of C as in its absence), p_c is the *generative causal power* (i.e., capacity or strength; Cheng, 1997) of C to produce E, estimated based on the event probabilities:

$$p_c = \frac{P(e = 1|c = 1) - P(e = 1|c = 0)}{1 - P(e = 1|c = 0)} \tag{2}$$

For our experimental variables, let R represent the outcome (the presence of an allergic reaction, as indicated by whether red dots appeared on a patient's face), let p_a, p_b, p_x, and p_y respectively represent the generative causal powers (Cheng, 1997) of binary cues A, B, X, and Y to produce R, and let a, b, x, y, $r \in \{0, 1\}$ denote the absence and presence of cues A, B, X, and Y, and outcome R. With respect to our blocking design, the noisy-OR causal-invariance function assumes that cue X's causal strength is invariant regardless of whether cue A is present. Likewise, the function assumes that cues B and Y's causal strengths are invariant across contexts. In accord with causal invariance, by Eq. (2) the generative power of cue X (the answer to the rating question regarding the consumption of food X by itself) is:

$$p_x = \frac{P(r = 1|x = 1) - P(r = 1|x = 0)}{1 - P(r = 1|x = 0)} = \frac{1 - 1}{1 - 1} = \frac{0}{0},$$

which has an undefined value, whereas the generative power of cue Y is:

$$p_y = \frac{P(r = 1|y = 1) - P(r = 1|y = 0)}{1 - P(r = 1|y = 0)} = \frac{0 - 0}{1 - 0} = \frac{0}{1} = 0.$$

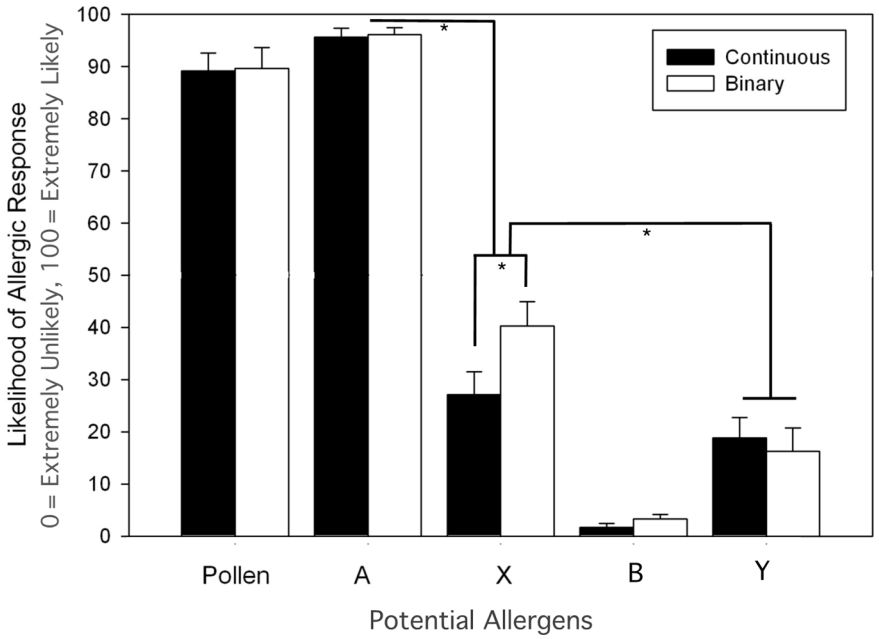

Fig. 2. Mean ratings of how likely the patient would develop an allergic reaction following exposure to or consumption of each of the potential allergens. Error bars represent one standard error of the mean. An asterisk (*) indicates a statistically significant difference, with p ranging from 0.025 to 0.045.

Because the undefined value of the causal strength of cue X is potentially greater than 0, but control cue Y has a causal strength of 0, cue X should receive a higher causal-strength rating than cue Y. That is, cue X should be only "incompletely blocked" by cue A. Lu et al.'s (2008) Bayesian version of causal strength assuming Eq. (2) and uniform priors makes the same predictions as just explained.

The blocking design thus provides a qualitative test of the appropriate use of the two quantitative causal-invariance integration functions. Comparing across the predictions for the two conditions, we see that cue X should receive a lower causal-strength rating in the Continuous condition than in the Binary condition. Moreover, the difference between the ratings for cues X and Y should be smaller in the Continuous condition than in the Binary condition. Finally, because cue A should be similarly perceived as having a high causal strength in both conditions, the difference between the ratings for cues A and X should be larger in the Continuous condition than the Binary condition. All three of these predicted between-group differences indicate a bigger blocking effect for the Continuous than the Binary condition.

Note that in the more typical blocking design – with A+, AX+ in the Experimental condition and BY+ in the control condition – cues B and Y are controls for X (see our description earlier). This design, unlike ours, does not have any control cue with a low estimated causal strength as a baseline against which to demonstrate the predicted higher estimated causal strength of X when the outcome is binary: for all three cues, X, B, and Y, there would be high uncertainty. Under the typical design, the pattern of ratings in support of the noisy-OR causal-invariance function over the additive function would be a null result: namely, "no difference" in the Binary condition between the ratings for cue X and control cues B and Y. The support would involve the pitfall of accepting the null hypothesis. Moreover, ratings for control cues B and Y may show high variability due to the confounding; neither cue occurs by itself. Our design therefore offers a more sensitive test for differentiating between our Continuous and Binary conditions, and hence a more sensitive test of the causal-invariance view against alternative views.

3.2.3. Results

Participants' mean ratings for the likely occurrence of an allergic reaction following each cue are depicted in Fig. 2. As the figure shows, participants' ratings varied substantially between cues. However, the ratings for the cues across the two groups were quite similar with the exception of cue X, the cue of interest. In particular, the similar ratings between groups for ragweed pollen and cue A confirm that the maximal outcome level was perceived similarly across groups.

To examine the overall pattern of results, a 2 (Condition: Continuous vs. Binary, between-subjects) by 4 (Cue: A, B, X vs. Y, within-subjects) mixed analysis of variance was conducted. Because Mauchly's test of sphericity indicated a significant violation, $\chi 2(5) = 56.41$, $p < .001$, degrees of freedom were corrected (lowering them) using Greenhouse-Geisser ($\varepsilon = 0.63$). The analysis revealed a significant main effect of cue, $F(1.9, 87.4) = 331.6$, $MSE = 243.3$, $p < .001$, $\eta_p^2 = 0.878$. There was not a significant main effect of condition, $F(1, 46) = 1.79$, $MSE = 268.8$, $p = .19$, $\eta_p^2 = 0.037$, consistent with the cues in general (including our control cues) being rated similarly across conditions. Neither was there a significant overall interaction between cue (including controls) and condition, $F(1.9, 87.4) = 2.33$, $MSE = 243.3$, $p = .11$, $\eta_p^2 = 0.048$.

Recall that only the causal invariance view predicts that participants who receive continuous pre-training will produce stronger "blocking" of critical cue X than those who receive binary pre-training. Specifically, this prediction means that, comparing causal-strength ratings between the two groups, 1) cue X would receive lower causal-strength ratings from participants in the Continuous group, 2) the difference in ratings for cue X and the control cue Y would be smaller in the Continuous group, and 3) the difference in ratings for X and the control cue A would be larger in the Continuous group. Because these three predictions are all directional, we conducted one-tailed planned comparisons for each hypothesis using Welch's independent t-tests with $\alpha = 0.05$ (unless otherwise noted; 95% CIs for cell means were always computed two-tailed).

In support of our first prediction, we found that participants who received continuous pre-training gave significantly lower ratings for critical cue X ($M = 27.08$, $SD = 21.67$, 95% CI = [17.96, 36.21]) than participants in the Binary group ($M = 40.21$, $SD = 23.29$, 95% CI = [30.40, 50.02]), $t(45.8) = 2.02$, $p = .025$, $d_s = 0.58$. This difference can be seen between the two groups' mean ratings in the X column of Fig. 2. (All other cues were predicted by all accounts to be equal between groups; all two-tailed t's less than 1.46, p's greater than 0.15.)

To further evaluate the lower mean causal-strength rating of cue X in the Continuous than in the Binary group, our next two hypothesis tests examine cue-by-condition interactions, comparing across groups the difference scores for participants' causal-strength ratings for critical cue X compared to their ratings for control cue Y and for control cue A, respectively. Recall that higher ratings for cue X than cue Y indicate incomplete blocking of cue X, a prediction made by the causal-invariance view for the Binary group alone. Confirming this predicted interaction, the difference between the ratings for cues X and Y was significantly larger in the Binary group ($M = 23.96$, $SD = 34.67$, 95% CI = [9.35, 38.57]) than the Continuous group ($M = 8.29$, $SD = 27.33$, 95% CI = [-3.22, 19.81]), $t(43.6) = 1.74$, $p = .045$, $d_s = 0.50$, as can be seen from comparing the X and Y columns of Fig. 2. Corroborating this observed interaction, one-tailed dependent t-tests for each group revealed that the effect of pre-training on the X-Y difference scores was driven by the Binary group alone demonstrating incomplete blocking—the rating for cue X in the Binary group ($M = 40.21$, $SD = 23.29$, 95% CI = [30.40, 50.01]) was substantially and significantly higher than that for control cue Y ($M = 16.25$, $SD = 21.82$, 95% CI = [7.05, 25.45]), $t(23) = 3.38$, $p = .001$, $d_z = 0.69$. By contrast, the rating for cue X in the Continuous group ($M = 27.08$, $SD = 21.67$, 95% CI = [17.96, 36.21]) was not significantly higher than that for control cue Y ($M = 18.79$, $SD = 19.03$, 95% CI = [10.78, 26.81]), $t(23) = 1.49$, $p = .075$, $d_z = 0.30$, a pattern consistent with greater blocking.

In support of our second hypothesized interaction, the differences between the ratings for cues A and X were significantly larger in the Continuous group ($M = 68.54$, $SD = 22.82$, 95% CI = [58.93, 78.15]) than the Binary group ($M = 55.83$, $SD = 21.70$, 95% CI = [46.69, 64.98]), $t(45.9) = 1.98$, $p = .027$, $d_s = 0.57$, as can be seen from comparing the A and X columns of Fig. 2.

In summary, all three planned comparisons converge to suggest that reasoners who received continuous pre-training produced stronger blocking, and those who received binary pre-training produced incomplete blocking. All between-group comparisons yielded Cohen's d_s effect sizes of 0.5 or greater. Note that whereas these observed differences between groups were predicted *a priori* by the causal-invariance view, none is explicable by any alternative views (e.g., Beckers et al., 2005; Rescorla & Wagner, 1972), including those based on empirically learned causal-integration functions (e.g., Griffiths & Tenenbaum, 2009; Lucas & Griffiths, 2010; Melchers et al., 2004; Mitchell et al., 2009).

3.2.4. Discussion

Experiment 1 tested our hypothesis that reasoners have analytic knowledge of the respective causal-invariance functions for binary and continuous outcome variables—knowledge essential for guiding search in the infinite space of representations. We manipulated participants' interpretation of an outcome as binary or continuous without conveying information on how causal influences combine, while keeping all trials in the blocking paradigm and all other information identical across conditions. The information held constant included domain-specific knowledge of integration functions, the maximal intensity of the outcome, and the range of the intensity of the outcome. Our results show that participants who received information conveying that the outcome was continuous rather than binary showed a greater blocking effect, confirming all three differences predicted by our hypothesis. Empirically learned causal-integration functions—acquired either before or during our study—predict no difference between conditions, because all experience involving cue combinations was kept constant across randomly-assigned conditions. Similarly, the empirical-verification explanation (e.g., Beckers, et al., 2005, 2006; De Houwer et al., 2005; Lovibond et al., 2003; Mitchell et al., 2009; Mitchell & Lovibond, 2002) predicts no blocking for either condition: because the outcome during blocking training occurred at the maximal observed level for all participants, none had information available to verify that adding a second causal cue would have raised the level of the outcome. Our results provide support for our causal invariance hypothesis.

A logical implication of the empirical-inferential account (e.g., Beckers et al., 2005; Mitchell et al., 2009) is that the blocking effect should not occur in the standard blocking design. In the standard "A+, AX+" blocking design (Kamin, 1968), the outcome occurs at a single intensity whenever it does occur (e.g., the experimenter sets a shock to either occur or not) — the outcome therefore occurs at the maximal experienced strength in the A+ phase and there is no information for verifying that a higher intensity can occur. However, the blocking effect has been demonstrated across multiple species (e.g., Couvillon, Lianne Arakaki, & Bitterman, 1997; Merchant & Moore, 1973; Rodrigo, Chamizo, McLaren, & Mackintosh, 1997; Sahley, Rudy, & Gelperin, 1981). These findings are inexplicable by the empirical-inferential account, but consistent with our causal-invariance account if the outcome variable type is perceived to be continuous. We return to this issue in the general discussion.

A result we had not expected was that the mean rating for control cue Y was clearly higher than that for cue B, the noncausal control cue. The difference was observed in both conditions, $t(23) \geq 2.87$, $p < .01$; for the Continuous condition, $M = 17.13$, $SD = 19.63$, 95% CI = [9.27, 24.98], and for the Binary condition, $M = 12.92$, $SD = 21.21$, 95% CI = [4.43, 21.40]. To demonstrate our predicted differences, the unexpectedly high rating for cue Y requires that Cue X in the Binary condition show a more robustly higher rating than would have been necessary otherwise. This requirement would have been a problem for our conclusion if it had led to a failure to obtain a) the predicted higher rating for cue X than cue Y in the Binary Condition, or b) a significantly larger difference in mean rating between cue X and cue Y in the Binary condition than in the Continuous condition. However, as reported, we did observe both predicted first- and second-order differences. Our conclusion therefore remains fully valid.

The non-zero rating for cue Y in our B−, BY− condition may reflect uncertainty about its causal status. Because cue Y, like cue X, was never presented alone, uncertainty about its causal status exists. Multiple possibilities might have come to a participant's mind, for example, Cues B and Y interact, Cue B is preventive, etc.

Cue Y serves well as a control cue for X because its rating represents a baseline uncertainty about a cue only observed as part of a compound. Notably, neither our theory nor any of the other causal-learning theories in the psychological literature predicts our observed rating for Y. Our participants, being students in a psychology course, might have been especially sensitive to this uncertainty due to their common training on "no confounding" as a pre-requisite for causal inference. What is crucial, again, is that overlaying the inherent uncertainty that likely contributed to the ratings for both cues X and Y, the difference between the ratings for X and Y was nonetheless significantly greater in the Binary than in the Continuous condition—an interaction predicted by the causal-invariance account alone.

3.2.4.1. Relation to previous studies. Our paper tests the causal-invariance hypothesis qualitatively, but the quantitative predictions of the causal invariance function for a binary outcome have been supported by other studies across multiple labs (see Lu et al. 2008 for a review and additional discriminating tests). Simulation findings from Lu et al.'s (2016) sequential Bayesian modeling of experimental results from previous human and nonhuman blocking studies suggest that participants assumed additivity or the noisy-OR, respectively, depending on whether the outcome variable was continuous or binary. Our results are consistent with these findings, and go further by providing an explanation for what unites the mathematically different functions conceptually. Because the simulations are based on studies that varied in multiple ways—in participants' species and prior experiences, experimental tasks, instructions, and other factors—the findings do not warrant the interpretation that these integration functions were used in their role as causal invariance. By solely manipulating the perceived outcome as binary or continuous within a species, Experiment 1 provides the first empirical evidence for the causal-invariance interpretation of those functions.

The work supporting the empirical-inferential reasoning account manipulated participants' empirical knowledge of how specific causes combined their influences. Domain-specific knowledge was conveyed variously by prior familiarity with a particular domain (e.

g., Blanco, Baeyens, & Beckers, 2014), verbal instructions on how causes combine their influences on the same outcome (e.g., Mitchell & Lovibond, 2002), or a pre-training phase in which participants received different exposures to how causes in a domain or context combined their influences (e.g., Beckers et al., 2005, 2006; Lovibond et al., 2003). Previous studies never manipulated perceived outcome-variable type while keeping all experiences constant.[6] Because manipulations of empirical knowledge of integration functions fully explain participants' knowledge of these functions across conditions, these studies have not tested whether reasoners have analytic knowledge of causal-invariance functions.

A set of experiments by Beckers et al. (2005, Experiment 1; Beckers et al., 2006, Experiment 3) illustrates the support for their empirical-inferential account. The experiments manipulated whether the highest intensity of the outcome during a pre-exposure phase was higher than the outcome during subsequent blocking training. The outcome during blocking training was presented either at the maximal level experienced during pre-exposure or at a submaximal level. Beckers et al. explain the greater blocking effect observed in the submaximal group by this group's prior experienced higher level of the outcome, enabling empirical verification of the failure of cue X to increase the level of the outcome over that produced by cue A alone.

More generally, all studies by Beckers et al. (2005, 2006), Blanco et al. (2014), Collins and Shanks (2006), and Lovibond et al. (2003) showed no difference between conditions in the blocking effect unless there was either a manipulation of 1) experienced outcome maximality, or 2) empirical knowledge regarding the integration function prior to blocking training. In contrast, our experiment obtained a difference in the blocking effect between conditions without either manipulation. The observed difference is instead explained purely by the disambiguation of the continuous or binary nature of the outcome variable, and hence the respective analytic knowledge of causal-invariance integration functions. As explained in Section 2.2, prior empirical learning cannot account for the adoption of these respective functions.

Turning to a different aspect of previous related studies, we note that they have often focused on the additivity assumption. Beckers et al. (2006, p. 98), for example, considered additivity as a default assumption, noting that it "might … explain why many contemporary associative models, in which additivity is a hardwired feature of associative learning … have been so successful in explaining the bulk of conditioning phenomena." Researchers tested additivity (e.g., Collins & Shanks, 2006; Soto et al., 2009) in elemental associative models (e.g., Rescorla & Wagner, 1972; Van Hamme & Wasserman, 1994) against configural models (e.g., Pearce, 1987, 1994), or made use of pre-training on additivity or subadditivity to challenge associative models (Beckers et al., 2006; De Houwer, Beckers, & Vandorpe, 2005, Lovibond et al, 2003).

As a result of the focus on additivity, the discussion has often been in terms of methodological constraints that might have precluded support for the additivity assumption. For example, Collins and Shanks (2006, p. 1533) write, "it appears that those previous findings [against additivity] may have been a function of the fact that the outcome values assigned to the stimuli were binary and maximal, and participants were unable to give a magnitude response." Their studies eliminated several of these potentially constraining factors and showed support for additivity. Likewise, Lovibond et al. (2003, p. 134) write, "the design of most laboratory causal judgment tasks imposes … a ceiling on effect magnitude, because they present the effect as binary". The binary nature of the outcome seems to be viewed as a methodological shortcoming.

Our framework shifts the focus of discussion from additivity *per se* to the more general concept of causal invariance, and introduces a distinction between whole causes and their interacting components. In our framework, additivity should not be adopted as a default at the whole-cause level for binary outcomes, because it does not denote causal invariance for such outcome variables. However, we do agree with previous researchers that—in the context of a blocking paradigm involving deterministic causal relations, as is typical and as was the case in Experiment 1— the binary nature of the outcome does impose a methodological constraint, as we explain presently.

3.2.4.2. Support for causal invariance using the blocking paradigm across and within conditions in Experiment 1. The greater blocking effect observed in our Continuous-outcome than Binary-outcome condition is inexplicable by the use of one single integration function for both outcome variable types. For a continuous outcome, the event pattern "A+, AX+" is consistent with A being causal and X being noncausal, with additivity as the invariance integration function. It is inconsistent with X being causal. In contrast, for a binary outcome, the event pattern "A+, AX+" allows for the possibility that both cues are causal, and they combine their strengths in an invariant manner as described by Eq. 1.

However, the event pattern for a binary outcome is also consistent with a positive causal interaction between two cues. This ambiguity stems from two features of the blocking paradigm: 1) it does not independently vary two stimulus cues (i.e., X never appears alone), and 2) its event patterns are deterministic. Given these features, a participant who judged cue X to be causal on its own may nonetheless assume that the two cues interact. A positive interaction in this case would not further boost the probability of the binary outcome, because that probability was already 1.0 in our deterministic design. Although explaining "A+, AX+" assuming an interaction seems unlikely when causal invariance provides a simpler explanation, the more complex explanation cannot be ruled out. In other words, while our results are consistent with participants in our Continuous and Binary conditions adopting the appropriate causal-invariance function for the respective outcome-variable types, it is logically possible that while some participants in the Binary condition defaulted to assuming the invariance of two causal cues according to Eq. 1, others defaulted to assuming a positive

[6] Lu et al (2016) erroneously reported that Mitchell and Lovibond (2002) gave participants instructions indicating a binary or a continuous outcome. Mitchell and Lovibond in fact conveyed that the outcome was a continuous variable for both their "additive" and "non-additive" conditions: At the beginning of their experiments prior to the manipulations, all participants were instructed to vary the intensity level of the shock they would receive (the unconditioned stimulus) to select a level that was "definitely uncomfortable, but not painful" to them (p. 314–315). The adjustment of intensity levels should have conveyed to all participants that the outcome variable was continuous.

interaction between two causes of a binary outcome (Novick & Cheng, 2004).

The deterministic feature of our blocking design is necessary due to our experimental-design requirement to hold the blocking trials constant across the Continuous and Binary conditions. The strength of a cue to produce an outcome is signified differently for the two outcome types: it is signified by the probability of a binary outcome versus the magnitude of a continuous outcome. The only situation in which the blocking trials in our two conditions can be kept identical (for the purpose of testing people's spontaneous interpretation) is when the cue-outcome pairings are deterministic: a cue is either always paired with an outcome occurring with a constant magnitude (7 dots on the face) or never paired with it.

To disambiguate support for the assumption of causal invariance as the default for a binary outcome, Experiment 2 presented participants with a scenario involving two independently varied candidate causes of a binary outcome, with the causes producing the outcome probabilistically in one condition and deterministically in another. In the probabilistic condition, causal invariance and deviation from it would be distinguishable. Experiment 2 thus complements our test of causal invariance in Experiment 1.

3.3. Experiment 2

Our primary goal in Experiment 2 is to provide converging evidence to disambiguate the support from Experiment 1 for the default status of causal invariance (rather than causal interaction) for binary outcomes in hypothesis revision. To serve this goal, Experiment 2 employed the overexpectation paradigm involving two independently varied candidate causes of a binary outcome (Kremer, 1978; Rescorla & Wagner, 1972), extending it by manipulating whether the causal relations in question are deterministic or probabilistic. When two probabilistic causes are independently varied, the requisite information for judging whether or not the causes interact with each other is available. Judgment of interaction is precluded in the blocking paradigm, in which one of the two candidate causes in question (the blocked cue, X) never occurs alone.

Our secondary goal is to demonstrate the causal nature of the causal-invariance assumption, specifically, to demonstrate that the causal-invariance function is used in a causal rather than merely associative manner. Findings on the use of integration functions to combine multiple influences have often been explained in causal terms. The empirical-inferential reasoning account, for example, is framed causally (e.g., Beckers, et al., 2005; Blanco et al., 2014; De Houwer et al., 2002; Lovibond et al., 2003; Mitchell et al., 2009; Mitchell & Lovibond, 2002), as seen in the passage in Beckers et al. quoted earlier. It is possible, however, that reasoners in fact do not interpret integration functions causally, but instead use them purely as mathematical functions. Moreover, the verification argument in this account applies to mere association as well as causation. For example, two effects (e_1 and e_2) of a common cause C are associated with each other, but merely so since neither effect causes the other. Now, if another cause D occurs on every trial and produces e_2 at a maximal level (the analog of A+ trials in the blocking paradigm), the association between e_1 (the analog of cue X) and e_2 (the analog of +) would be impossible to verify. Regardless of whether e_1 occurs, e_2 occurs.

Participants in Experiment 2 were given outcome-frequency information on various (fictitious) proteins called "endomins" in blood serum and asked to judge whether each variant of these proteins causes hypertension. We assessed the causal nature of the integration function in three related ways: whether participants judged the causal status of these proteins by 1) choosing subsets of events that honor the "no confounding" condition, a prerequisite for causal inference, 2) using the additive integration function without heeding the "ceiling effect" (e.g., as does the Rescorla-Wagner model, 1972), and 3) revising their hypothesis due to a deviation from additivity rather than from causal invariance.

Our extension of the overexpectation paradigm with a deterministic (ceiling) versus probabilistic (non-ceiling) manipulation enables the pursuit of a third goal: to show that a binary outcome need not be an obstacle to demonstrating qualitative additivity (e.g., Beckers et al., 2005, 2006), contrary to what sometimes seems to have been assumed (e.g., Lovibond et al., 2003; Collins & Shanks, 2006; Soto, Vogel, Castillo, & Wagner, 2009). Because a binary outcome by definition occurs at a maximum level whenever it occurs, it is tempting to treat such outcomes as imposing a ceiling condition. But Eq. 1 specifies varying probabilities of a binary outcome over the entire range from 0 to 1.0, and only the extreme values (i.e., causal determinism) impose a ceiling condition. Thus, as we explain presently, analytic knowledge of Eq. 1 predicts whether or not the phenomenon of overexpectation occurs depending on whether the cues in the paradigm produce the binary outcome at a *ceiling* or *non-ceiling* level (i.e., with a probability of 1.0 or less, respectively). Crucially, causal invariance and additivity coincide in their qualitative predictions in the non-ceiling case (both predict overexpectation), but not in the ceiling case (causal invariance predicts no overexpectation).

Fig. 3 schematically summarizes the stimulus–response pairings in the standard (deterministic) overexpectation paradigm, which we adopted for our *ceiling* situation (Experiment 2B). In the figure, the letters A and B represent stimulus cues, and a bar above A or B indicates the absence of the cue. The paradigm has two phases. The design for Phase 1 is represented in the top contingency tables in the figure, for the Experimental and Control groups, respectively. The Experimental group's stimulus pattern is A+, B+: namely, when either cue A or B is present while the other is absent, the outcome always occurs (indicated by a "+" in the corresponding cell in the table); otherwise, when both cues are absent, the outcome does not occur (indicated by an "O"). Cues A and B do not simultaneously occur in Phase 1 (as indicated by the empty cell). For the Control group (see the upper-right table), the design is A−, B+ . The presence of Cue A by itself is paired with the absence of the outcome (see upper right cell). This is the only difference from the Experimental group.

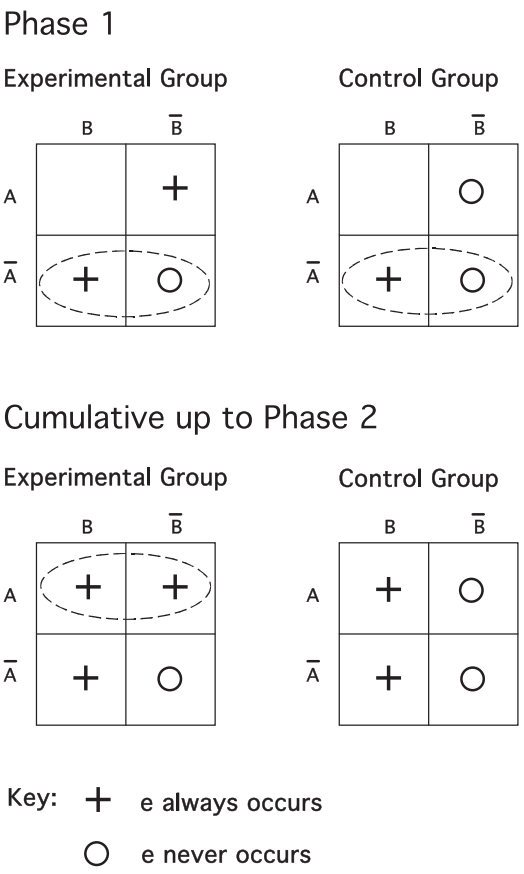

Fig. 3. Contingency tables representing the occurrence of effect *e* in the overexpectation design. The letters A and B represent stimulus cues, and a bar above a letter indicates the absence of the cue. A dashed ellipse in a table encloses the subset of events that satisfies the "no confounding" pre-requisite for computing the causal power of critical cue B; the universal set violates this pre-requisite in all except the bottom right table.

To facilitate explaining the various model predictions, we constructed the bottom row of tables in Fig. 3 to represent the cumulative stimulus-outcome pairings across Phases 1 and 2. (In other words, the bottom tables repeat the pairings from Phase 1 along with the additional pairings in Phase 2.) In Phase 2, the design is AB+ for both the experimental and control groups: namely, when both cues A and B occur at the same time, the outcome always occurs, and when neither cue occurs, the outcome does not occur (represented by the "+" and "O" cells in the diagonal of the bottom contingency tables).

The phenomenon of *overexpectation*—a reduction in the estimated causal strengths of the elemental cues from Phase 1 to Phase 2 for the Experimental group, compared to the Control group—has been observed with this design in animal conditioning studies (Kremer, 1978; Rescorla & Wagner, 1972) and in human causal learning (Collins & Shanks, 2006) when the causal relations are deterministic.

Experiments 2A and 2B presented events fitting the overexpectation paradigm in the non-ceiling and ceiling situations, respectively. Experiment 2B adhered to the deterministic paradigm just described. Experiment 2A involved a *non-ceiling* situation in which the outcome occurs—in those cells with "+"s in the figure —with a constant intermediate probability that is clearly less than 1 and clearly greater than 0. (That is, Fig. 3 describes the design for this situation if the "+"s now indicate the intermediate probability.) If Cues A and B are individually shown to be probabilistic causes of the outcome in Phase 1, then the expected probability of the outcome given the combination of A and B in Phase 2 can be inferred assuming their individual causal powers remain invariant in the context of the other cause. This enables an assessment of a deviation of the observed probability from the expected probability. The ceiling situation in Experiment 2B (determinism) precludes such an assessment.

Below we first report the experimental method for the two situations, then explain our various model predictions. Our key predictions concern participants' revision at the end of Phase 2 of their initial hypotheses at the end of Phase 1, depending on their assessment of whether there is a deviation from invariance across the two phases. Some readers are likely to find our results for the two situations intuitively obvious. We tested the situations nonetheless because the difference in responses across situations provides a discriminating test of the causal invariance hypothesis against alternative accounts.

3.3.1. Experiments 2A and 2B: Non-ceiling and ceiling situations

3.3.1.1. Method

3.3.1.1.1. Materials, design, and procedure. For both experiments, college students were presented with a cover story in which they were asked to infer what proportion of patients who have various "newly discovered" (fictitious) proteins called "endomins" in their bodies would have hypertension. Trials consisted of individual patients. Hypertension, the outcome in question, was presented as a binary variable.[7]

The procedure was identical across experiments. In each experiment, participants were randomly assigned to either an Experimental or a Control group. They were presented with two successive learning phases, with the information presented on a desktop computer. On each learning trial, participants saw a "hospital record" listing an individual patient's initials, along with information about the presence or absence of three endomins in the patient. Some patients' bodies produce some of the endomins, and different patients' bodies produce different patterns of endomins. On each trial, participants pressed a button to indicate whether they thought the patient with that record had hypertension, after which they received feedback on their prediction and whether the patient indeed had hypertension, along with their cumulative prediction accuracy.

We adapted the standard overexpectation design to enable a comparison between our ceiling and non-ceiling situations. In Phase 1, the patients displayed four types of endomin patterns: R only, S only, T only, or no endomins. Endomins R and S map onto cues A and B respectively in Fig. 3; endomin T is a manipulation check and not represented in the figure (we explain its role presently). Because all relevant comparisons are between groups, the labels of the endomins were not counterbalanced. In Phase 2, the patients displayed three types of patterns: the RS compound (i.e., R and S together), T only, or no endomins. Tables 1 and 2 list the presented proportions of patients with hypertension for each endomin pattern in each phase for the Experimental and Control groups of Experiments 2A and 2B, respectively. For both experiments, within each phase, participants viewed records of 24 patients with each of the endomin patterns in that phase, interleaved within a block in a different random order for each participant.

For Experiment 2A (as shown in Table 1), in the Experimental group the proportion of patients with hypertension in Phase 1 was 0.75 for both endomins R and S individually. Endomin T has a higher proportion, 0.92, and was added to the standard overexpectation paradigm to provide a manipulation check confirming that a probability of 0.75 was perceptibly below a (near) ceiling level. In Phase 2, the proportion was 0.75 for the RS compound and 0.92 for T. None of the patients without endomins had hypertension in either phase.

For the Control group, the sole difference was that Endomin R alone was not paired with hypertension in Phase 1. Recall that Endomin R maps onto cue A in Fig. 3. Note that Endomin S (cue B in the figure) in the Control group is the critical cue for comparison with Endomin R or S in the Experimental group.

For Experiment 2B, the materials, design, and procedure were identical to those of Experiment 2A, with the exception of the conditional probabilities of hypertension (see Table 2): Specifically, the proportion of patients with hypertension was 1 in Experiment 2B whenever it was non-zero in Experiment 2A.

After each block in both phases, all participants were asked to judge the relative frequencies of hypertension for each of 5 patterns of endomins: the three individual endomins, the RS compound, and no endomins. Participants were asked, "Out of 100 patients with this pattern of endomins, how many do you think have hypertension?" Their answers for the individual endomins are estimates of the elemental causal strengths of the endomins. Their answers for the RS compound in the two phases allow us to evaluate whether participants did expect a higher frequency for the compound based on their observations in Phase 1 (before the compound had appeared), compared to their estimate based on observations in Phase 2. A higher expectation for the RS compound would confirm that the deviation from this expectation was indeed the reason for a reduction in the causal estimates of the individual cues R and S from Phase 1 to Phase 2.

To collect data efficiently, the computer program monitoring the experiment stopped for participants who gave unusually poor estimates of the relative frequencies of the 4 presented stimulus patterns in Phase 1: Specifically, Phase 2 proceeded only when the participant's estimates were within 20 points (out of 100) of the actual proportions, and an estimate for T that was at least one point higher than those for R alone and S alone. Participants were given a maximum of 3 blocks of trials to reach the criterial accuracy.

To reduce unnecessary noise in the data, our cover story encouraged one interpretation of the source of any perceived conflict between phases: Every participant in both experiments was told at the start of Phase 2 that there may or may not have been some inaccurate diagnoses of patients whose records they viewed in Phase 1, but that the diagnoses of the patients whose records they were going to see from then on were certainly accurate. They were told that if unexpected trends appeared, there must have been some errors in the earlier records, but that if there were no unexpected trends, the earlier records must have been all correct. Importantly, our instructions left it entirely to the participants to judge whether or not there was a deviation from expectation.

3.3.1.2. Model predictions. We derive predictions based on four dominant learning models, with tests of the hypothesis – that causal invariance is the default integration function – as our goal and focus. (Model fitting is not our goal.) Below we first explain the predictions according to the causal-invariance integration function for binary outcomes for the non-ceiling and ceiling situations. Toward our goal of demonstrating the causal nature of this integration function, we contrast this function's predictions with those according to three non-causal models: 1) an elemental associative model that uses additivity as the mathematical integration function regardless of

[7] Although the cover story technically involved an observational study, the wording in our story and our questions encouraged a causal interpretation of the data.

Table 1
Actual Proportion of Cases (Out of 24) with Hypertension and Participants' Estimated Proportion of Cases (Out of 100) with Hypertension in Each Phase for Experiment 2A (Non-ceiling Situation). Standard deviations for the estimated number of patients out of 100 are enclosed in parentheses; 95% confidence intervals for the estimated number of patients out of 100 are enclosed in square brackets.

	Phase 1				Phase 2			
	Experimental Group		Control Group		Experimental Group		Control Group	
	Actual	Estimated	Actual	Estimated	Actual	Estimated	Actual	Estimated
R	.75	.76 (10.5) [71.2, 80.8]	0	.01 (.02) [0.08, 1.92]	—	.68 (16.9) [60.2, 75.8]	—	.02 (.09) [1.96, 2.04]
S	.75	.78 (9.5) [73.6, 82.4]	.75	.74 (13.0) [68.1, 80.0]	—	.68 (16.4) [60.5, 75.6]	—	.72 (9.5) [67.6, 76.4]
T	.92	.90 (8.4) [86.1, 93.9]	.92	.92 (8.3) [88.2, 95.8]	.92	.90 (9.5) [85.6, 94.4]	.92	.89 (10.6) [84.1, 93.9]
R&S	—	.90 (9.3) [85.7, 94.3]	—	.70 (13.6) [63.7, 76.3]	75	.70 (19.1) [61.2, 78.8]	.75	.71 (12.4) [65.3, 76.7]
none	0	.00 (0) [.00, .00]	0	.00 (0)[.00, .00]	0	.00 (0) [.00, .00]	0	.00 (0) [.00, .00]

Table 2
Actual Proportion of Cases (Out of 24) with Hypertension and Participants' Estimated Proportion of Cases (Out of 100) with Hypertension in Each Phase for Experiment 2B (Ceiling Situation). Standard deviations for the estimated number of patients out of 100 are enclosed in parentheses; 95% confidence intervals for the estimated proportions are in square brackets.

	Phase 1				Phase 2			
	Experimental Group		Control Group		Experimental Group		Control Group	
	Actual	Estimated	Actual	Estimated	Actual	Estimated	Actual	Estimated
R	1.0	1.0 (0) [100.0, 100.0]	0	.01 (2.4) [-0.5, 2.5]	—	1.0 (0) [100.0, 100.0]	—	.0 (0) [0.00, 0.00]
S	1.0	.99 (2.4) [97.5, 100.5]	1.0	.99 (2.4) [97.5, 100.5]	—	1.0 (0) [100.0, 100.0]	—	1.0 (0) [100.0, 100.0]
T	1.0	1.0 (0) [100.0, 100.0]	1.0	1.0 (0) [100.0, 100.0]	1.0	1.0 (0) [100.0, 100.0]	1.0	1.0 (0) [100.0, 100.0]
R&S	—	1.0 (0) [100.0, 100.0]	—	1.0 (0) [100.0, 100.0]	1.0	1.0 (0) [100.0, 100.0]	1.0	1.0 (0) [100.0, 100.0]
none	0	.0 (0) [0.0, 0.0]	0	.0 (0) [0.00, 0.00]	0	.0 (0) [0.00, 0.00]	0	.0 (0) [0.00, 0.00]

its interpretation as causal invariance (e.g., as in Rescorla & Wagner, 1972), 2) a configural associative model that generalizes based on similarity (Pearce, 1987, 1994), and 3) a contingency model (Jenkins & Ward, 1965), which estimates associative strength based on the entire set of events presented, rather than on a subset that satisfies the "no confounding" prerequisite.

Prediction of overexpectation (a reduction in causal-strength estimates in Phase 2) by every model depend on whether there is a deviation from the expected outcome across the two phases—a conflict that warrants a resolution. But, only the expected outcome according to the causal view predicts the intuitively obvious pattern: there should be overexpectation when the endomins individually are only sometimes paired with the binary outcome (the *non-ceiling* situation), but not when they are always paired with the outcome (the *ceiling* situation). As will become clear, such a pattern of hypothesis revision is inconsistent with assuming an interaction between stimulus cues as the default.

3.3.1.2.1. Causal invariance. The causal power theory (Cheng, 1997) explains why *only* when a candidate cause is "not confounded" (e.g., when alternative causes are held constant) is the generative causal power (i.e., capacity or strength) of candidate cause C to produce effect E inferable, as specified in Eq. (2). The theory assumes causal invariance (Eq. 1). Tables 1 and 2 list the predicted causal strengths according to the theory. Subsets of events in which the "no confounding" prerequisite holds, and hence Eq. (2) applies, are termed *focal sets* (Cheng and Novick, 1992). Dashed lines in Fig. 3 circle the focal sets for estimating the strength of critical cue B.

3.3.1.2.1.1. Predictions for the non-ceiling situation

Consider the Experimental group. In Phase 1 of this group (see top-left contingency table in Fig. 3), for cue B, the only focal set is the circled set in the contingency table, where the only alternative cause is the constantly present "background" (cue A is absent in this set). Applying Eq. (2) to this focal set yields a causal power of 0.75 for cue B. The same value results for cue A (in both the ceiling and non-ceiling situations, cues A and B in the Experimental group are symmetrical). We also see that—under the assumption that A and B each retains their causal power regardless of whether the other cue is present (Eq. 1) — the estimate of the proportion of patients with hypertension resulting from the AB compound (unseen in this phase) should be higher than estimates for either cue alone. Specifically, an invariance-assumed estimate for the AB compound is $0.75 + 0.75 - (0.75 \times 0.75) = 0.9375$ (Eq. 1).

In Phase 2 (recall that the design was AB+), for the Experimental group, the only focal set for cue B in this phase is one that accumulates information across the two phases (the circled set in the bottom left table of Fig. 3), in which cue A is constantly present. In this set, according to Eq. (2), $p_B = 0$; that is, cue B's causal power is now 0. The same logic applies to cue A. Therefore, for both cues, an inconsistency should appear between their causal powers inferred in the two phases (0.75 versus 0). This does not occur for the Control group[8], for whom consistently across phases cue B's causal power is 0.75 and cue A's causal power is 0.

Recall that for all participants, in between Phases 1 and 2, our cover story conveyed that the observations in Phase 1 may contain

[8] For the Control group in Phase 2, because cue A has been established non-causal in Phase 1, the state of this cue makes no difference to the causal assessment of cue B in Phase 2. Thus, there is no need to select a subset of events for the assessment of cue B.

inaccuracies but that observations in Phase 2 are to be trusted. Returning to the Experimental group, under this cover story, for a reasoner in the non-ceiling condition who assumes causal invariance, reducing the estimate of the causal powers of the two cues from 0.75 to 0.50 would resolve the apparent inconsistency across the two phases (because by Eq. 1, $0.5 + 0.5 - (0.5 \times 0.5) = 0.75$). This resolution implies that participants in the Experimental group, but not those in the Control group, should show overexpectation, that is, give lower estimates of hypertension in patients with cue B (Endomin S) after seeing Phase 2.

Regarding our manipulation check, observing a higher causal rating for Endomin T alone than Endomin S alone would confirm that Endomin S is correctly perceived to occur at a non-ceiling level.

3.3.1.2.1.2. Predictions for the ceiling situation

The causal-invariance view predicts no overexpectation in the ceiling situation (the "+"s in Fig. 3 now indicate hypertension occurring with a probability of 1). For the Experimental group in Phase 2, cues A and B each has an undefined causal-power value due to $p_A = p_B = 0/0$ (Eq. (2)). Therefore, continuing to believe that these cues have the same causal power as in Phase 1 does not result in any conflict, and participants should have no reason to change their causal assessment of the cues. For the Control group, consistently across Phases 1 and 2, $p_A = 0$ and $p_B = 1$, also resulting in no deviation from invariance. Hence, there should be no overexpectation in either group in the ceiling situation.

In summary, finding the just-explained difference in overexpectation between the ceiling and non-ceiling situations would support the causal-invariance hypothesis. Because our cover story about endomins was identical for all participants, prior knowledge of the domain cannot account for the difference.

3.3.1.2.2. Rescorla and Wagner's (1972) model. Here we explain the predictions according to the additive integration function of Rescorla and Wagner's (1972) model. Given that the outcome occurs with the same probability across the two phases, in the ceiling as well as in the non-ceiling situation, the arithmetic sum of the associative strengths of cues A and B after Phase 1 "overpredicts" the outcome associated with the AB compound in Phase 2 in both situations. This deviation from expectation in Phase 2, according to this model, should result in overexpectation for both cues A and B, for both the ceiling and non-ceiling situations.

3.3.1.2.3. Pearce's (1987) model. The associative learning rule incorporated in Pearce's (1987) model differs from that in Rescorla and Wagner's model in two ways. First, the layer of nodes that serves as input to the learning rule consists of configural units (the third layer in Pearce's network) rather than elemental units representing the stimuli. Each stimulus pattern, including those that consist of a single cue, is represented by its respective configural unit. Second, on each trial, the associative strength of only one configural unit—the one representing the cue combination present—will be updated.

Applying this model to the overexpectation design, we see that the configural units respectively representing A alone and B alone will not have their associative strengths updated in Phase 2, because these stimuli are not presented alone in this phase. The lack of strength revision implies there will be no strength reduction in these units. In other words, unlike Rescorla and Wagner's (1972) model, Pearce's (1987) model predicts no overexpectation for either cue A or cue B, for either the ceiling or non-ceiling situation.

3.3.1.2.4. The "no confounding" prerequisite often absent in associative models. "No confounding" is not a prerequisite for noncausal learning models. A reasoner who does not seek the "no confounding" condition may default to using the entire set of events presented. We make use of the *ΔP model* (Jenkins & Ward, 1965), which treats the universal set as input, to test predictions for the omission of this prerequisite. This model predicts causal strength by ΔP, the difference in the probability of effect E in the universal set when candidate C occurs and when C does not occur:

$$\Delta P = P(e = 1|c = 1) - P(e = 1|c = 0) \tag{3}$$

First, note that for the Experimental group (the left contingency tables in Fig. 3), the ΔP value for neither cue A nor cue B changes across Phases 1 and 2 for either the ceiling or non-ceiling situation. Accordingly, this model predicts that reasoners should perceive no conflict across phases. Thus, like Pearce's (1987) model, ΔP predicts that reasoners should show no overexpectation for either situation.

A second consequence of omitting the "no confounding" prerequisite concerns the strength estimate for cue B in Phase 1 (see top tables in Fig. 3). In this phase, the ΔP value for cue B is lower in the Experimental than in the Control group. This is because $P(e = 1|b = 0)$ is higher for the Experimental group (in the *absence* of cue B in this group, cue A is paired with the outcome). Thus, cue B should receive a *lower* estimate in the Experimental group than in the Control group. This prediction holds for both the ceiling and non-ceiling situations. In contrast, honoring the "no confounding" prerequisite yields the prediction that the causal power of cue B in Phase 1 should be the same across the two groups (see circled sets in the top contingency tables) for both situations.

3.3.2. Experiment 2A: The Non-ceiling situation

3.3.2.1. Method

3.3.2.1.1. Participants. Based on the variability in a pilot study, we *a priori* chose to collect data until 20 participants complete the experiment in each of the two conditions. Forty-seven undergraduates at the University of California, Los Angeles (UCLA) were recruited to participate in the experiment to satisfy a course requirement. None of the participants had been exposed to formal probability theory. Four and 3 participants, respectively, from the Experimental and Control groups did not reach the criterial accuracy described in 3.3.1.1.1 and no test-phase data were collected from them. After excluding these participants, there were 20 participants in each group.

The design, materials, and procedure were as explained earlier.

3.3.2.2. Results and discussion. Table 1 presents the results of Experiment 2A. The table also lists the actual proportions of patients with the various endomin patterns who have hypertension presented in each phase. Participants' mean causal estimates shown in the table — the estimated proportions of patients with given endomin patterns who have hypertension after the last block of each phase — are in accord with the predictions of both the causal power theory (1997) and Rescorla and Wagner's (1972) model, but contradict the predictions of Pearce's (1987) model and the ΔP model (Jenkins & Ward, 1965).

A 2 (Group: Experimental vs Control; between-subjects) by 2 (Phase: Phase 1 vs Phase 2; within-subjects) by 5 (Evaluation Cue Pattern: Endomin R, S, T, Endomins R + S, no endomin; within-subjects) mixed analysis of variance shows the three-way interaction predicted by causal invariance and Rescorla and Wagner's model, $F(4,152) = 10.5$, $MSE = 44.8$, $p < .0001$, $\eta^2 = 0.22$. The reduction of causal estimates for particular cue patterns across phases differed across groups. There were main effects of Group (Experimental $M =$ 63.0 vs. Control $M = 47.0$), $F(1,38) = 80.6$, $MSE = 318.4$, $p < .001$, $\eta^2 = 0.68$; Phase (57.1 vs 52.8), $F(1,38) = 13$, $MSE = 141.1$, $p <$ .001, $\eta^2 = 0.26$; and Cue Pattern (35.8, 72.8, 90.7, 75.5, vs 0), $F(4,152) = 992.3$, $MSE = 108.9$, $p < .001$, $\eta^2 = 0.96$. There were also two-way interactions between Group and Phase, Group and Cue Pattern, and Phase and Cue Pattern, $F(1,38) = 9.1$, $MSE = 141.1$, $p =$ .005, $\eta^2 = 0.20$, $F(4,152) = 179.3$, $MSE = 108.9$, $p < .001$, $\eta^2 = 0.83$, and $F(4,152) = 6.8$, $MSE = 44.8$, $p < .001$, $\eta^2 = 0.15$, respectively.

To further examine the pattern of results, planned one-tailed *t*-tests were conducted for all predicted directional differences. Recall that, for participants in the Experimental group, the causal invariance view and Rescorla and Wagner's (1972) model both predict that 1) estimate for the unseen RS (AB) compound during Phase 1 should be higher than for R or S alone, and 2) individual estimates for R (A) and S (B) should decrease in Phase 2 (i.e., overexpectation should occur). Moreover, this decrease in the individual estimates for R and S across phases in the Experimental group should be greater than that in the Control group.

The observed causal estimates support these predictions. First, in the Experimental group at the end of Phase 1, the mean estimated proportion with hypertension for the unseen RS compound (0.90) was significantly higher than that for R and S individually, $t(19) =$ 5.72 and 4.77, respectively, $p < .001$ for each, $d = 1.28$ and 1.07 respectively, in accord with use of either the additivity or noisy-OR function. Moreover, confirming that the probability of hypertension given the presence of R or S alone was perceived to be below ceiling, the estimated proportion for either endomin R (0.76) or S (0.78) was significantly lower than that for endomin T (0.90), $t(19)$ = 5.58 and 4.69, respectively, $p < .001$ for both, $d = 1.25$ and 1.05, respectively. Critically, the pattern of estimated proportions obtained after Phase 2 confirmed the phenomenon of overexpectation: A reduction in the estimate for the critical endomin S relative to Phase 1 was observed in the Experimental group. In this group, the mean estimate for endomin S significantly decreased by 0.10 (from 0.78 to 0.68) across the two phases, $t(19) = 4.75$, $p < .001$, $d = 1.06$. The corresponding reduction in the Control group was 0.02 and not statistically significant, $t(19) = 0.97$, $p = .34$. Moreover, the mean reduction in the estimate for S in the Experimental group was significantly greater than the corresponding difference observed in the Control group, $t(38) = 2.56$, $p = .005$, $d = 0.81$, consistent with the Group by Phase interaction.[9]

Analyzing the number of participants who did or did not show a reduction in their estimate for cue S indicates the same pattern of results. Fifteen participants in the Experimental group showed a decline in their estimate for S from Phase 1 to Phase 2, and 5 did not (of these, 4 gave equal estimates in the two phases and 1 showed an increase). The reverse pattern occurred in the Control group: 5 participants showed such a decline, and 15 did not (of these, 10 gave equal estimates and 5 showed an increase), and the frequency of decline differed across groups, X^2 (1, N = 40) = 10.0, $p = 0016$. To summarize, more participants in the Experimental group reduced their estimates, in line with the causal-invariance predictions and Rescorla and Wagner's (1972) model.

Recall that Pearce's (1987) model and Jenkins and Ward's (1965) contingency model predict no overexpectation. The observed pattern of overexpectation disconfirms both models and supports the causal nature of reasoners' inferences: reasoners honor the "no confounding" prerequisite for causal inference. Our participants' responses indicate that they spontaneously selected subsets of events in which alternative causes occur independently of the candidate cause.

A second result corroborates the honoring of this prerequisite. Despite the lower unconditional contrast for cue S in the Experimental than Control group, the groups did not significantly differ in their estimates of S in Phase 1 (0.78 vs 0.74 respectively), $t(38) =$ 1.41, $p = .45$. There is therefore no support for a noncausal contingency interpretation of the event frequencies.

In conclusion, our results obtained in Experiment 2A strongly favor both the causal power theory (1997) and Rescorla and Wagner's (1972) model over Pearce's (1987) and Jenkins and Ward's (1965) models. We now turn to the ceiling situation to test predictions that distinguish the causal invariance view from Rescorla and Wagner's.

3.3.3. Experiment 2B: Ceiling situation

Recall that for the ceiling situation, the causal invariance view predicts no overexpectation (participants should perceive no conflict in estimated causal power across the two phases), whereas Rescorla and Wagner's (1972) model continues to predict overexpectation.

3.3.3.1. Method

3.3.3.1.1. Participants. Based on the variability in a pilot study, we determined *a priori* to collect data from 24 participants

[9] The same pattern of results obtained using the average of the individual R and S estimates for each participant rather than their estimate for S alone in our analysis. We note that although the reduction is different between groups, the mean reduction in the Experimental group is smaller than what would be required to fully resolve the conflict if participants remembered the experimental instructions at the beginning of Phase 2 about the possible inaccuracy of the initial set of medical records (the fully invariant estimate would be a reduction to 0.5). The discrepancy could be due to 1) some participants' failure to remember the instructions by the end of Phase 2, 2) inaccurate estimates of the deviation from expectation, or 3) a tendency for early evidence to influence causal judgments more than does later evidence (Marsh & Ahn, 2006).

randomly assigned to the Experimental and Control conditions. Twenty-five UCLA undergraduates served in Experiment 2B to satisfy a course requirement, one more than planned due to experimenter error. They were randomly assigned to two groups, 12 to the Experimental group and 13 to the Control group. No participant was excluded on the basis of the stopping criteria.

3.3.3.1.2. Materials, design, and procedure. Experiment 2B presented participants with a ceiling situation, in which hypertension occurs with a probability of 1 whenever it occurs (i.e., causal determinism). The materials, design, and procedure were as explained earlier.

3.3.3.2. Results. Table 2 presents the actual and estimated proportions of patients with hypertension for the Experimental and Control groups. In contrast to the non-ceiling situation tested in Experiment 2A, participants' estimates for the critical endomin S in Experiment 2B were unchanged across phases (ranging from 0.99 to 1.0) in both the Experimental and Control groups. All differences between groups were far from statistical significance, two-tailed $t(23) \leq 0.64$, $p \geq 0.52$ for all comparisons between groups. The number of participants who did or did not show a reduction in their estimate for endomin S corroborates the absence of overexpectation. None of the participants in either the Experimental or the Control group showed a reduced estimate for endomin S: 11 participants in each group gave equal estimates across phases, and respectively 1 and 2 participants in the Experimental and Control groups showed a small increase (by less than 0.09).

At the end of Phase 1, consistent with the absence of overexpectation, and notably as predicted by reasoners' use of analytic knowledge of binary outcomes, the Experimental group estimated that the proportion for the unseen RS compound was 1.0 (not different from the group's estimates for the individual endomins), rather than greater than 1.0. The latter is analytically impossible (by reason, there cannot be more patients with hypertension than there are patients), but it is predicted by the additivity function in Rescorla and Wagner (1972).

In conclusion, in support of the predictions of the causal invariance view and Pearce's (1987) model, but contradicting those of Rescorla and Wagner's (1972) model, our results from both phases in a ceiling situation indicate that the Experimental group did not show overexpectation.

Comparing across Experiments 2A and 2B, the frequency of participants showing (vs. not showing) a reduced estimate for critical endomin S in Phase 2 relative to Phase 1 of the Experimental condition was greatly and significantly larger in Experiment 2A than in Experiment 2B, X^2 (1, $N = 32$) = 16.9, $p < .0001$. Corroborating this difference predicted by the causal invariance view alone, an analysis of variance across the two experiments shows the expected 3-way (Experiment by Group by Phase) interaction for endomin S, F (1, 76) = 4.6, $p = .035$, $\eta^2 = 0.055$. The two experiments were conducted under highly similar conditions.

3.3.3.3. Discussion. In summary, using an otherwise identical experimental design, we observed overexpectation in a non-ceiling situation (Experiment 2A) but not a ceiling situation (Experiment 2B). This pattern of results is as predicted by the causal-invariance function for binary outcome variables (Cheng, 1997), but not by explanations based on either the use of an additive integration function (Rescorla & Wagner, 1972) or the use of a contingency criterion without satisfying the "no confounding" prerequisite for causal inference (Jenkins & Ward, 1965). A configural associative model that generalizes based on similarity (Pearce, 1987; 1994) likewise fails to explain the pattern. The inferential reasoning account (e.g., Beckers et al., 2006; Mitchell et al., 2009) does not explain why participants' causal-strength estimates were based on spontaneously selected subsets of events that satisfied the "no confounding" prerequisite. Neither does it explain previous quantitative findings in support of use of the noisy-OR and noisy-AND-NOT causal-invariance functions for a binary outcome (e.g., for a review, see Lu et al., 2008). In addition to clarifying the causal and analytic nature of reasoners' use of a causal-invariance function, our findings demonstrate that an outcome being binary is not an obstacle to testing whether people detect a deviation from additivity. Finally, by ruling out an assumption of interaction between causes of a binary outcome that was left ambiguous in Experiment 1, Experiments 2A and 2B add converging support for the causal-invariance hypothesis.

4. General discussion

4.1. Summary

Our paper addresses the age-old challenge of understanding the causal-induction problem that biological intelligent agents evolved to solve: how generalizable causal knowledge is possible when our only access to reality is via our observations in terms of represented entities and relationships. It explains why causal invariance across the learning and application contexts is a necessary aspiration in a rational agent's construction of a useable causal representation of the world.

In addition to explicating this computational-level question (Marr, 1982) about the goal of causal learning (Cheng & Lu, 2017), our paper presents converging empirical support for causal invariance as an aspiration. In particular, our experiments tested whether intuitive causal learning in humans is equipped with analytic knowledge of the respective causal-invariance functions for two outcome variable types common in everyday encounters. In Experiment 1, we found that—while holding constant participants' domain knowledge of how causes combine their influences, blocking paradigm trials, and the experienced maximal outcome—merely encouraging the interpretation of the outcome variable as either continuous or binary was sufficient to invoke the appropriate causal-invariance function for combining multiple causal influences for the two outcome-variable types. Our exclusion of experience of integration functions, both prior to and during our experiment, as a possible explanation for our results argues for the availability of analytic knowledge of the causal-invariance functions for these outcome-variable types.

In Experiments 2A and 2B, the pattern of cue-competition effects on a binary outcome involving two independently varied stimulus cues provides converging evidence for the causal-invariance explanation, over the causal-interaction explanation, of the reduced blocking of critical cue X in the Binary condition of Experiment 1. The pattern of effects also highlights the causal nature of the integration function for binary outcomes.

Together, our findings offer initial evidence favoring our hypothesis—that people use their analytic knowledge of causal invariance to guide their causal hypothesis formation and revision—over alternative accounts of causal induction that have not addressed the challenge of this problem in the infinite search space of representations (e.g., Beckers et al., 2005; De Houwer et al., 2002, 2005; Gopnik et al., 2004; Jenkins & Ward, 1965; Lucas et al., 2014; Melchers et al., 2004; Mitchell et al., 2009; Pearce, 1994; Pearl, 2000; Rescorla and Wagner, 1972; Spirtes et al., 1993/2000). Our findings are what would be expected in view of the representation-dependent reality in the human mind. Causal invariance as a tacit aspiration gives a unified explanation of both qualitative findings (Experiments 1, 2A, & 2B; Liljeholm & Cheng, 2007; Wu & Cheng, 1999) and quantitative findings (e.g., Buehner et al., 2003; for a review, see Lu et al., 2008) testing the use of decomposition functions.

In the remaining sections, we discuss five related issues: 1) the implication of our findings for explaining the puzzle of why the blocking effect sometimes occurred and other times not, in seemingly highly similar situations, 2) the implicit nature of analytic knowledge of causal invariance, 3) the relation between such knowledge and development work on invariances (Karmiloff-Smith, 1992; Piaget, 1950, 1985), 4) the apparent tension between causal invariance as an aspiration and as a description (e.g., Griffiths & Tenenbaum, 2005; 2009; Lucas & Griffiths, 2010; Melchers et al., 2004; Wheeler et al., 2008; Urcelay & Miller, 2010), and 5) the dependency of invariance judgments on representation in visual perception (Cheng & Pachella, 1984).

4.2. *A potential explanation of the elusiveness of the blocking effect in rodents (Maes, Boddez, Alfei, Krypotos, D'Hooge, De Houwer, & Beckers, 2016)*

The robustness of the blocking effect has been challenged by a group of researchers who reported failures to replicate the effect in 15 rodent studies despite the use of procedures that are highly similar or identical to those in published studies (Maes et al., 2016). Maes et al. conclude, "those failures raise doubts regarding the canonical nature of the blocking effect and call for a reevaluation of the central status of blocking in theories of learning" (p. 1).

While recent emphases on replicability have unearthed many serious problems with reported psychological findings (Open Science Collaboration, 2015), our conjecture is that in this particular case, the apparent elusiveness of the blocking effect may in part be explained by the ambiguity of the outcome-variable type in the standard blocking training. In the typical paradigm, the unconditioned stimulus (US, e.g., an electric shock at a certain amperage) occurs at two levels: it either occurs or does not occur. There are two plausible interpretations of such an outcome. One is of course that the outcome is binary, and the only two possible values are the ones observed. A second is that the particular intensity level of the US that was experienced is incidental—merely a selected value along a continuum of possible values. This interpretation is viable in that the typical US causes pain, relieves hunger, or quenches thirst, and the animal has likely experienced these outcomes at varying levels of intensity prior to the experiment—due to similar causes that themselves may have varied in intensity. For example, a laboratory rat has bumped into obstacles that may have caused various intensities of pain, and background causes that affect the perceived intensity may have varied (e.g., variations in an experimenter's handling). Thus, although the experimental outcome is designed to be binary by the researcher (the shock either does or does not occur), it may nonetheless be perceived by the animal as continuous.

Experiment 1 may be regarded as a test of our explanation of the elusiveness of the blocking effect. When the outcome-variable type of the experimental materials is ambiguous, our causal invariance hypothesis implies that different causal-invariance functions will be used depending on a reasoner's interpretation of the outcome variable. As a result, the blocking effect will appear or not appear, seemingly capriciously. Although our experiment was conducted on humans and cannot be directly compared with rodent studies, its results showing differing amounts of blocking as a function of the perceived outcome variable type are consistent with our explanation of the elusive blocking effect.

4.3. *Relation to related developmental work (Karmiloff-Smith, 1992; Piaget, 1950, 1985)*

Our analytic concept of causal invariance is complementary to the empirical concept of invariance discussed in developmental work (Karmiloff-Smith, 1992; Piaget, 1950, 1985). At a general level, these concepts share the core assumption that cognition has a goal of developing a stable representation of some aspect of the world that generalizes across variation in other aspects. Piaget's (1950) use of the concept of invariance concerns the empirically acquired invariances of physical properties, and Karmiloff-Smith's (1992) use concerns re-descriptions in children's development of their understanding of the physical world, progressing from an implicit and less generalizable representation of a stable knowledge state to an explicit and more generalizable one.

The shared core of the concept bolsters the basic premise of our work. Our concept differs from these developmental variants, however, in two complementary respects. First, both Piaget's (1950) and Karmiloff-Smith's (1992) concepts of invariance concern the resulting outputs, intermediate or final, of domain-specific *empirical invariances*, such as the invariance of the property of specific physical materials (e.g., the invariance of the volume of a liquid when it is poured from a wide glass into a narrow glass) or of an empirically acquired invariant law (e.g., the law characterizing how a balance beam can be made to stay horizontal by objects of different weights positioned on the two sides of the beam across a fulcrum). While their findings provide evidence that young children already show an aspiration toward obtaining invariant representations, such domain-specific knowledge are the outputs of the causal-learning process. In contrast, our analytic version of the concept is a domain-independent navigating device within the learning

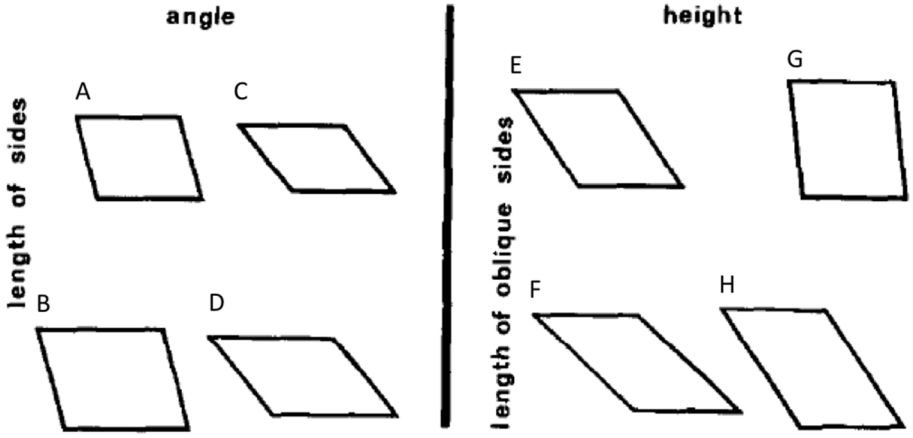

Fig. 4. Left panel: Parallelograms varying orthogonally on angle and length of sides (with ratio of lengths of adjacent sides kept constant). Right panel: Parallelograms varying orthogonally on height and length of oblique sides (with length of horizontal sides kept constant).
Source: Cheng and Pachella (1984, p. 284)

process, directing the search toward those outputs. For children as for adults, what gets revised is the domain-specific knowledge. Causal invariance as a revision criterion and an aspiration is not revised.

Second, the developmental processes characterized in Karmiloff-Smith (1992) and Piaget's (1950) work differ both from each other and from ours. Karmiloff-Smith's (1992) representational-redescription model describes "a process of 'appropriating' stable states to extract the information they contain, which can then be used more flexibly for other purposes" (p. 25). Notably, re-description occurs *only* when a stable successful state is achieved. It concerns cognitive changes in terms of re-descriptions of the same content state (e.g., the stable final understanding of balancing objects on a beam) at different levels of explicitness and abstraction. The re-description changes how explicit that understanding is, without changing its content. In contrast, the re-representation predicted by our concept of causal invariance occurs when the knowledge state is unstable, when there is a deviation from invariance, and does involve changing content (e.g., our fiber and metabolic syndrome example; the particle representation of photons versus the wave-particle duality of photons; see discussion in Cheng & Lu, 2017). These apparently opposite conditions of re-representation in fact do not contradict each other, because the accounts are about disparate aspects of invariance despite use of the same linguistic term. (Our next section addresses the implicit versus explicit nature of knowledge of the causal invariance function for a binary outcome.)

Causal-knowledge revision due to deviation from analytic causal invariance may be regarded as a specific form of Piaget's (1985) process of equilibration, in which knowledge revision towards equilibrium takes place when the system is in a state of disequilibrium. Equilibrium is a state of satisfaction, and disequilibrium is a state of discomfort and confusion that occurs when an individual's current understandings are challenged, through new information or experiences. For example, disequilibrium may occur when a child's belief that lions do not climb trees is corrected by a teacher. Piaget's "disequilibrium" is a more general concept than deviation from causal invariance in that it need not concern causal knowledge. Our work extends Piaget's in that: 1) we specify the mathematical criterion by which disequilibrium for causal knowledge would result, and 2) our framework explains why children, like scientists, aspire to construct representations that are stable and invariant.

4.4. Implicit knowledge of the causal invariance function for a binary outcome

People's knowledge of the invariance functions for a binary outcome variable is likely to be in an implicit form only. In view of the challenge and the importance of adaptive causal induction, it seems unlikely that evolution would have left the solution up to the explicit reasoning of individual members of a species. People may not have explicit knowledge of the noisy-OR function because it is not typically taught in school. To measure whether people have explicit knowledge of the causal-invariance function for a binary outcome, we tested 31 UCLA undergraduate students from the same participant pool as our experiments on their use of the noisy-OR function in an explicit non-causal form. The participants were told that an urn has 8 blue dice and 2 red dice. A single die is randomly chosen from the urn and rolled. The die is then returned to the urn and the process repeats. When participants were asked about the probability of a conjunctive event, "What is the probability of picking a red dice *and* rolling a '4'?", they all gave the correct answer (1/30). But when they were asked about the probability of a disjunctive event, "What is the probability of picking a red dice *or* rolling a '4'?", only 10% answered this question correctly (the correct answer is 1/3).[10]

The low level of performance for the non-causal "urn" version of the noisy-OR question is in stark contrast to the implicit use of the noisy-logicals in a causal setting. In numerous causal learning experiments (e.g., Buehner et al., 2003; Liljeholm & Cheng, 2007; Lu et al., 2008), the majority of participants estimated causal strength or judged causal structure as if they had correctly computed the

[10] No participant gave an answer (3/10) consistent with the "exclusive OR".

probability of the analogous disjunctive event (a target outcome produced by cause A *or* by cause B). Thus, it seems that the analytic knowledge of the noisy-OR function used by participants in these causal-induction experiments is only in an implicit form. It is, of course, possible to explicitly teach the noisy-OR formula, but reasoners appear not to need to know the explicit form in order to reason with that analytic knowledge to estimate causal strength.

4.5. Whether or not features are invariant depends on the chosen representation: A perceptual example

For a concrete example of how invariance is dependent on representation, let us consider the description of a parallelogram, with our percept as the target outcome (Cheng & Pachella, 1984). Parallelograms do not in themselves possess parameters that combine invariantly or interactively. A parallelogram can be specified by any three orthogonal parameters, such as the lengths of two adjacent sides and an angle, the lengths of one side and one diagonal and the height, and so on. These possible representations do not all produce perceptually invariant features. To keep our illustration simple, let us vary only two dimensions by fixing the third.

The left panel of Fig. 4 depicts four parallelograms varying orthogonally along two dimensions: the acute angle and length of the sides, with the ratio of the lengths of adjacent sides fixed. This representation illustrates that shape is perceptually invariant across variations in size. As can be seen, despite variation in the size of the parallelogram due to variation in the length of the sides, parallelograms with the same acute angle (compare A and B) are geometrically 'similar'—they are invariant in shape, a perceptually salient feature.

Now, what if we orthogonally vary the height of the parallelogram and the length of the oblique sides, with the length of the horizontal sides fixed? As illustrated in the right panel, parallelograms of a constant height (compare G and H) vary in both shape and size, as do parallelograms of a constant length of the oblique sides (compare E and G), depending on the value on the other dimension. In other words, these two physical dimensions interact psychologically. As Fig. 4 illustrates, whether two dimensions are "separable" or "integral" depends on the variables in the chosen representation. Returning to the causal analogs of this visual example, note that across domains, perceptual and conceptual, invariance or non-invariance across context depends on the reasoner's constructed representations of reality.

4.6. Resolving the roles of causal invariance as aspiration and as description in complex causes with multiple interacting causal factors

The reader may have perceived a gap between causal invariance as an aspiration and as a description. After all, causes are often complex. Causal factors may interact in many possible ways, characterized by various integration functions. Treating causal invariance as an ideal may thus seem simplistic or wishful. Perhaps in part for this reason, the concept of causal invariance is not regarded as essential in the psychological literature, in Bayesian models (e.g., Griffiths & Tenenbaum, 2005; 2009; Lucas & Griffiths, 2010) or in propositional inference models (e.g., Beckers et al., 2006; Mitchell et al., 2009). Nor is it heeded in current mainstream statistics for testing causal hypotheses involving binary outcomes (e.g., Fienberg, 1980/2007; McCullagh & Nelder, 1989; Wickens, 1989).

The description side of the apparent gap is well illustrated by Griffiths and Tenenbaum's (2005) view on the appropriateness of an integration function. They write (p. 349), "We suspect that the appropriate parameterization for the relationship between a cause and its effects [i.e., the appropriate integration function] will depend upon an individual's beliefs about the causal mechanism by which those effects are brought about. For some causal mechanisms, other parameterizations may be more appropriate than the noisy-OR." Similarly, it has been argued that learning integration functions that relate the theory under construction to data allows "learners to flexibly identify a specific functional form for causal relationships in a given setting, instead of assuming a fixed functional form for all causal relationships," (Lucas & Griffiths, 2010, p. 123), the fixed functional form being a causal-invariance function often assumed in the previous literature (e.g., Cheng, 1997). We of course agree that it is useful to learn descriptions of how specific types of causes combine their influences and to let that empirical knowledge guide the prediction of outcomes.

Our framework closes the apparent gap between invariance as aspiration and as description. In our view, invariant whole causes are what a reasoner implicitly aspires toward, and the relation between a whole cause and its component factors is the resulting description that (ideally) meets the aspiration. Newton's inverse square law is an exemplar of a whole cause aimed at invariance and defined by a description of interactive components. The descriptions are justified by experience, but they are not acquired in a purely bottom-up manner. Instead, they are guided by the top-down aspiration toward whole causes that combine with each other by causal-invariance functions that are justified by reason.

A great appeal of a probabilistic conception of causal invariance as an aspiration is that it offers an adaptive advantage for causal learning: it enables the powerful problem-solving strategy of *decomposition* (decomposition in a different sense as that in "decomposition function") to the learning of complex causes with multiple interacting factors. The probabilistic criterion of causal invariance, which generalizes the deterministic variant (e.g., Sloman, 2005), enables decomposing (in the problem-solving sense) the acquisition of a whole cause into more feasible incremental steps. Specifically, it enables the learning of a single component of a complex cause, even when the other components of the complex cause are unknown or missing—conditions that hinder learning under a deterministic criterion. Changing a single component within a "whole" cause can change the probability of the outcome. The resulting change justifies inferring the component as causal (assuming that other components of the whole cause occur at a constant non-zero rate and there is no confounding due to alternative causes, see Cheng, 2000).

Another appeal of use of the criterion of causal invariance to form whole causes is that it enables the computation of the combined effect of any configuration of whole causes (once their individual influences are known), without the need to have observed the configuration (Carroll & Cheng, 2010). In other words, it enables *compositionality*. For example, according to Newton's law, for any possible configuration of celestial bodies, the combined gravitational force on a celestial body x is the superposition of (i.e., the vector

sum of) the separate gravitational forces on *x* from all other celestial bodies in the configuration. Neptune's discovery, for example, was made possible by explaining the discrepancy between the observed orbit of Uranus and its orbit predicted based on the superposition of gravitational forces from the Sun and the other known planets. Thus, in multiple ways, analytic causal-invariance functions enable the learning and use of empirically-based integration functions.

To further understand the complementary nature of aspirations and descriptions, it may help to see some differences between an acquired integration function and a causal-invariance function. The former encodes domain-specific *empirical knowledge*, is subject to modification by further experience, and does not transfer to dissimilar domains (Lucas & Griffiths, 2010, Expt. 5; Melchers et al., 2004; Wheeler et al., 2008). Wheeler et al., for example, conducted a series of three experiments on rats investigating the boundaries for transfer of an acquired interaction-integration function across tasks. Their experiments tested the limits of the generalizability of an acquired "subadditive" integration function. They trained rats on a subadditive function in one learning task involving a continuous outcome variable, and measured transfer to a subsequent learning task involving the same outcome variable. For a continuous-outcome variable in the range where the scale is at least interval, additivity represents causal invariance and subadditivity indicates an interaction. They reported that the acquired subadditive function failed to generalize when the spatial or temporal characteristics of the environment or of the stimuli changed across the learning and transfer tasks. The generalization similar to that observed in a previous study (Beckers et al., 2006) also disappeared with a longer temporal delay between the two tasks. Based on these findings, Urcelay and Miller (2010) conclude that "prior experiences transfer to new experiences only when the two tasks share similar spatiotemporal contexts, at least in rats".

In contrast to empirical integration functions, causal-invariance functions encode formal knowledge that is neither limited to domains with similar perceptual or other content-specific features, nor subject to modification due to new data. Note that when there is deviation from the criterion of causal invariance in a data set, what is given up is only causal invariance as a description for the current representations of the current content. The aspiration of representing knowledge in terms of invariant whole causes and the causal invariance functions for each outcome-variable type are not given up.

5. Conclusion

We have argued that only by knowing what to expect assuming causal invariance can one judge deviations from the goal of causal invariance, and that to form that expectation one must have analytic knowledge of causal-invariance functions for outcome-variable types that are important for survival or for understanding. What if we humans do not have causal invariance as our goal in causal induction? The crucial role played by causal invariance in the construction of useable causal knowledge suggests that without it, we would be like Alice asking the Cheshire Cat for directions without knowing where she wants to go (Carroll, 1865/1998). Wouldn't you agree with the cat, that which way she ought to go "depends a good deal on where you want to get to" (p. 89)?

Declaration of Competing Interest

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

Acknowledgements

Jooyong Park was supported by Seoul National University Research Grant for Humanities and Social Sciences. Preparation of this paper was in part supported by NSF Grant DBS 9121298 to Cheng. A preliminary report of Experiment 1 was presented at the 54th Annual Meeting of the Psychonomics Society. We thank Chris Carroll, Clark Glymour, Kelly Goerdert, and George Sperling for helpful discussion and Ulrike Hahn for help on the clarity of our paper. We also thank Chris Carroll for his assistance in data collection.

References

Atlas, J. D. (2005). *Logic, meaning, and conversation: Semantical Underdeterminacy, implicature, and their interface*. Oxford: Oxford University Press.

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.

Blanco, F., Baeyens, F., & Beckers, T. (2014). Blocking in human causal learning is affected by outcome assumptions manipulated through causal structure. *Learning & Behavior, 42*(2), 185–199. https://doi.org/10.3758/s13420-014-0137-y

Botvinick, M., & Cohen, J. (1998). Rubber hands 'feel' touch that eyes see. *Nature, 391*, 756.

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*, 1119–1140.

Carroll, C. D., & Cheng, P. W. (2010). In *The induction of hidden causes: Causal mediation and violations of independent causal influence* (pp. 913–918). Austin, TX: Cognitive Science Society.

Carroll, L. (1865/1998). *Alice's Adventures in Wonderland.* Chicago, Illinois: Volume One Publishing.

Cartwright, N. (1989). *Nature's capacities and their measurement*. Oxford: Clarendon Press.

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

Cheng, P. W., Liljeholm, M., & Sandhofer, C. (2013). In *Logical consistency and objectivity in causal learning* (pp. 2034–2039). Austin, TX: Cognitive Science Society.

Cheng, P. W., Liljeholm, M., & Sandhofer, C. (2017). *Analytic Causal Knowledge for Constructing Useable Empirical Causal Knowledge: Two Experiments on Preschoolers*. (London, UK): 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* (pp. 65–84). Oxford, England: Oxford Univ Press.
Cheng, P. W., & Pachella, R. G. (1984). A psychophysical approach to dimensional separability. *Cognitive Psychology, 16*, 279–304.
Collins, D. J., & Shanks, D. R. (2006). Summation in causal learning: Elemental processing or configural generalization? *The Quarterly Journal of Experimental Psychology, 59*(9), 1524–1534. https://doi.org/10.1080/17470210600639389
Couvillon, P. A., Arakaki, L., & Bitterman, M. E. (1997). Intramodal blocking in honeybees. *Animal Learning & Behavior, 25*, 277–282.
De Houwer, J., Beckers, T., & Glautier, S. (2002). Outcome and cue properties modulate blocking. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 55A*, 965–985.
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.
Diamond, J. (1997). *Guns, germs, and steel: The fate of human societies*. New York: W.W. Norton & Company.
Feller, W. (1957/1968). *An introduction to probability theory and its applications* (3rd edition, Vol. 1). New York: John Wiley & Sons.
Fienberg, S. E. (1980/2007). *The analysis of cross-classified categorical data* (2nd edition). Cambridge, MA: MIT Press.
Glymour, C. (2001). *The mind's arrows: Bayes nets and graphical causal models in psychology*. Cambridge, MA: MIT Press.
Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes Nets. *Psychological Review, 111*(1), 3–32.
Good, I. J. (1961). A Causal Calculus (I). *British Journal for the Philosophy of Science, 11*(44), 305–318.
Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51*(4), 285–386.
Gopnik, A., Griffiths, T. L., & Lucas, C. G. (2015). When younger learners can be better (or at least more open-minded) than older ones. *Current Directions in Psychological Science, 24*(2), 87–92. https://doi.org/10.1177/0963721414556653
Griffiths, T., & Tenenbaum, J. (2009). Theory-based causal induction. *Psychological Review, 116*, 661–716.
Hawking, S., & Mlodinow, L. (2010). *The grand design*. New York: Bantam Books.
Hofer, H., Singer, B., & Williams, D. R. (2005). Different sensations from cones with the same photopigment. *Journal of Vision, 5*, 444–454.
Hume, D. (1739/1987). *A treatise of human nature* (2nd edition). Oxford: Clarendon Press.
Hume, D. (1748/1975). *An enquiry concerning human understanding and concerning the principles of morals*. In L. A. Selby-Bigge, & P. H. Nidditch (Eds.) (3rd edition). Oxford: Clarendon Press.
Ichien, N. & Cheng, P. W. (in press). 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*. London, UK: Bloomsbury Press.
Jayadevan, V., Michaux, A., Delp, E., & Pizlo, Z. (2017). 3-D shape recovery from real images using a symmetry prior. In *Proceedings of International Symposium on Electronic Imaging, Computational Imaging XV, Burlingame, CA (Society for Imaging Science and Technology, Springfield, VA, 2017)* (pp. 106–115).
Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge, England: Cambridge University Press.
Jenkins, H., & Ward, W. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs, 7*, 1–17.
Kamin, L. J. (1968). "Attention-like" processes in classical conditioning. In M. R. Jones (Ed.), *Miami symposium on the prediction of behavior: Aversive stimulation* (pp. 9–31). Miami, FL: University of Miami Press.
Kant, 1. (1781/1965). *Critique of pure reason*. London: Macmillan.
Karmiloff-Smith, A. (1992). *Beyond modularity: A developmental perspective on cognitive science*. Cambridge, MA: MIT Press.
Kremer, E. F. (1978). The Rescorla-Wagner model: Losses in associative strength in compound conditioned stimuli. *Journal of Experimental Psychology: Animal Behavior Processes, 4*, 22–36.
Kuhn, T. S. (1962/2012). *The structure of scientific revolutions* (50th anniversary edition). Chicago: University of Chicago Press.
Liljeholm, M., & Cheng, P. W. (2007). When is a cause the "same"? Coherent generalization across contexts. *Psychological Science, 18*, 1014–1021.
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., Rojas, R. R., Beckers, T., & Yuille, A. L. (2016). A Bayesian theory of sequential causal learning and abstract transfer. *Cognitive Science, 40*, 404–439.
Lu, H., Yuille, A. L., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review, 115*, 955–982.
Lucas, C. G., Bridgers, S., Griffiths, T., & 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.
Lucas, C. G., & Griffiths, T. L. (2010). Learning the form of causal relationships using hierarchical Bayesian models. *Cognitive Science, 34*, 113–147.
Ludwig, D. S. (2013). Examining the health effects of fructose. *Journal of American Medical Association, 310*(1), 34.
Luria, A. R. (1931). Psychological expedition to Central Asia. *Science, 74*, 383–384.
Mackie, J. L. (1974). *The cement of the universe: A study of causation*. Oxford, England: Clarendon Press.
Maes, E., Boddez, Y., Alfei, J. M., Krypotos, A.-M., D'Hooge, R., De Houwer, J., & Beckers, T. (2016). The elusive nature of the blocking effect: 15 failures to replicate. *Journal of Experimental Psychology: General, 145*, 1–23. https://doi.org/10.1037/xge0000200
Marr, D. (1982). *Vision*. New York: Freeman.
Marsh, J. K., & Ahn, W.-K. (2006). Order effects in contingency learning: The role of task complexity. *Memory & Cognition, 34*, 568–576.
McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd edition). Boca Raton: CRC Press.
Meder, B., Hagmayer, Y., & Waldmann, M. R. (2008). Inferring interventional predictions from observational learning data. *Psychonomic Bulletin & Review, 15*, 75–80.
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.
Merchant, H. G., III, & Moore, J. W. (1973). Blocking of the rabbit's conditioned nictitating membrane response in Kamin's two-stage paradigm. *Journal of Experimental Psychology, 101*, 155–158.
Miller, R. R., & Schachtman, T. R. (1985). Conditioning context as an associative baseline: Implications for response generation and the nature of conditioned inhibition. In R. R. Miller, & N. E. Spear (Eds.), *Information processing in animals: Conditioned inhibition* (pp. 51–88). Hillsdale, NJ: Erlbaum.
Mitchell, C. J., & Lovibond, P. F. (2002). Backward and forward blocking in human autonomic conditioning requires an assumption of outcome additivity. *Quarterly Journal of Experimental Psychology: Comparative & Physiological Psychology, 55B*, 311–329.
Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences, 32*, 183–198.
Mitchell, D. E., & Rushton, W. A. H. (1971). Visual pigment in dichromats. *Vision Research, 11*, 1033–1043.
Needham, J. (1956). *Science and civilisation in China: History of scientific thought* (Vol. 2). Cambridge, England: Cambridge University Press.
Needham, J. (1962). Science and civilisation in China. *Physics and physical technology* (Vol. 4). Cambridge, England: Cambridge University Press.
Needham, J. (1965). Science and civilisation in China. *Mechanical engineering* (Vol. 4.2). Cambridge, England: Cambridge University Press.
Needham, J. (1985a). Science and civilisation in China. *Paper and printing* (Vol. 5.1). Cambridge, England: Cambridge University Press.
Needham, J. (1985b). Science and civilisation in China. *Paper and printing* (Vol. 6.1). Cambridge, England: Cambridge University Press.
Needham, J. (1985c). Science and civilisation in China. *Fermentations and food science* (Vol. 6.5). Cambridge, England: Cambridge University Press.
Needham, J. (2000). *Science and Civilisation in China: Biology and Biological Technology, Part 6, Medicine*. Cambridge, England: Cambridge University Press.
Newton, I. (1687/1713). *The Principia: Mathematical Principles of Natural Philosophy*, translated and edited by I. Bernard Cohen and Anne Whitman, assisted by Julia Budenz (Berkeley and Los Angeles: University of California Press, 1999).
Novick, L. R., & Cheng, P. W. (2004). Assessing interactive causal influence. *Psychological Review, 111*, 455–485.

Park, J., & Sloman, S. A. (2013). Mechanistic beliefs determine adherence to the Markov property in causal reasoning. *Cognitive Psychology, 67*, 186–216. https://doi.org/10.1016/j.cogpsych.2013.09.002
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 Publishers.
Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge, England: Cambridge University Press.
Piaget, J. (1950). *The psychology of intelligence*. London: Routledge.
Piaget, J. (1985). *The Equilibration of Cognitive Structures: The Central Problem of Intellectual Development* (T. Brown & K. J. Thampy, Trans.). Chicago: University of Chicago Press.
Pizlo, Z. (2001). Perception viewed as an inverse problem. *Vision Research, 41*, 3145–3161.
Ramachandran, V. S., & Hirstein, W. (1998). The perception of phantom limbs: The D.O. *Hebb lecture. Brain, 121*, 1603–1630.
Rehder, B. (2014). Independence and dependence in human causal reasoning. *Cognitive Psychology, 72*. https://doi.org/10.1016/j.cogpsych.2014.02.002
Rehder, B. (2015). The role of functional form in causal-based categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(3), 670.
Rehder, B., & Burnett, R. C. (2005). Feature inference and the causal structure of categories. *Cognitive Psychology, 50*, 264–314. https://doi.org/10.1016/j.cogpsych.2004.09.002
Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black, & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York: Appleton-Century-Crofts.
Rodrigo, T., Chamizo, V. D., McLaren, I. P. L., & Mackintosh, N. J. (1997). Blocking in the spatial domain. *Journal of Experimental Psychology: Animal Behavior Processes, 23*, 110–118.
Rottman, B. M., & Hastie, R. (2014). Reasoning about causal relationships: Inferences on causal networks. *Psychological Bulletin, 140*, 109–139. https://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. https://doi.org/10.1016/j.cogpsych.2016.05.002
Sahley, C., Rudy, J., & Gelperin, A. (1981). An analysis of associative learning in a terrestrial mollusc. *Journal of Comparative Physiology. A, Neuroethology, Sensory, Neural, and Behavioral Physiology, 144*, 1–8.
Shanks, D. R., & Darby, R. J. (1998). Feature- and rule-based generalization in human associative learning. *Journal of Experimental Psychology: Animal Behavior Processes, 24*(4), 405–415. https://doi.org/10.1037/0097-7403.24.4.405
Sheps, M. C. (1958). Shall we count the living or the dead? *The New England Journal of Medicine, 259*, 1210–1214.
Sloman, S. (2005). *Causal models: How we think about the world and its alternatives*. New York: Oxford University Press.
Soto, F. A., Vogel, E. H., Castillo, R. D., & Wagner, A. R. (2009). Generality of the summation effect in human causal learning. *The Quarterly Journal of Experimental Psychology, 62*(5), 877–889. https://doi.org/10.1080/17470210802373688
Spirtes, P., Glymour, C., & Scheines, R. (1993/2000). *Causation, prediction and search* (2nd edition). Boston, MA: MIT Press.
Tenenbaum, J. B., & Griffiths, T. L. (2001). Structure learning in human causal induction. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *13. Advances in neural information processing systems* (pp. 59–65).
Urcelay, G. P., & Miller, R. R. (2010). On the generality and limits of abstraction in rats and humans. *Animal Cognition, 13*, 21–32.
Van Hamme, L. J., & Wasserman, E. A. (1994). Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. *Learning and Motivation, 25*, 127–151.
Wheeler, D. S., Beckers, T., & Miller, R. R. (2008). The effect of subadditive training 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: Oxford University Press.
Woodward, J. (2006). Sensitive and insensitive causation. *Philosophical Review, 115*, 1–50. https://doi.org/10.1215/00318108-2005-001
Woodward, J. (2010). Causation in biology: Stability, specificity, and the choice of levels of explanation. *Biology and Philosophy, 25*, 287–318.
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.