Causal Invariance Guides Inference of Empirical Integration Rules

Jeffrey K. Bye (jbye@umn.edu)
Department of Educational Psychology, 56 E. River Road
Minneapolis, MN 55455 USA

Patricia W. Cheng (cheng@lifesci.ucla.edu)
Department of Psychology, 1285 Franz Hall
Los Angeles, CA 90095 USA

Abstract

The present paper reports an experiment (*N*=254) testing two views of how reasoners learn and generalize potentially complex causal knowledge. Previous work has focused on reasoners' ability to learn *rules describing* how pre-defined candidate causes combine, potentially interactively, to produce an outcome in a domain. This *empirical-function learning* view predicts that reasoners would generalize an acquired combination rule based on similarity to stimuli they experienced in the domain. An alternative *causal-invariance* view goes beyond empirical learning: it allows for the possibility that one's current representation may not yield *useable* (i.e., invariant) causal knowledge — knowledge that *holds true when applied*. Accordingly, because useable causal knowledge is the evident aspiration of causal induction, this view posits that deviation from causal invariance is a criterion for knowledge revision. The criterion shapes the empirical functions learned and retained. A discriminating test is whether reasoners would re-represent interacting causes as a *whole cause* that does not interact with other causes, even when in their relevant experience all (pre-defined) causes in the domain interact. Our results favor the causal-invariance view.

Keywords: Causal induction; causal invariance; integration functions; empirical learning; analytic knowledge.

Introduction

In the final paragraph of Thomas Kuhn's (1962/2012) book, *The Structure of Scientific Revolutions*, he asks, "What must nature, including [we humans], be like in order that science be possible at all? … What must the world be like in order that [humans] may know it?" (p. 172). The cognitive-science side of that question is, "What must we humans *assume* the world be like, in order that we may know it?"

The present paper addresses this question with respect to the causal world. In particular, we experimentally test the *causal-invariance hypothesis* (Cheng & Lu, 2017), that reasoners a) assume there are such things as causes — potentially consisting of multiple interacting components — that produce effects (Kant, 1781/1965), and b) (implicitly) aspire to formulate knowledge such that causes are *invariant*, in other words, do not interact with other causes — and therefore this knowledge is "useable" in the sense that inferences from prior experiences hold in new situations with unknown other causes.

An example from science may serve to illustrate the hypothesis. For the reason that the process involved is basic to cognition itself, the causal-invariance hypothesis should apply to both intuitive and scientific reasoning. A causal-invariance assumption is embedded in Isaac Newton's (1687/1713) law of universal gravitation. The law states that the gravitational force between any two bodies is proportional to the product of their masses and inversely proportional to the square of the distance between them. For any pair of objects, the gravitational pull between them is governed by, and only by, this inverse-square law, *regardless of the existence and motion of all other objects*. This invariance property enables the motion of any object due to gravity to be predicted by the vector sum (the causal-invariance combination rule for vectors) of the gravitational pull on it from all other objects. Not all theories result in having this property that enables accurate prediction across different contexts, and theories that are more invariant are favored (Kuhn, 1962/2012; Woodward, 2000, 2010). Reasoners appear to aspire to maximize invariance. By *different contexts* with respect to an outcome, we mean situations that differ in the states of causes of the outcome that are present.

The causal-invariance hypothesis is an extension of causal learning models that make the *causal-invariance assumption* (e.g., Cheng, 1997; Lu et al., 2008). The assumption takes the form of a function that decomposes the observed outcome (e.g., the orbit of the Moon) into inherently unobservable contributions from all its causes (e.g., the Earth, the Sun, etc.), each with an *independent influence* on the outcome, an influence that does not change depending on the influences of other contributing causes. Functions that specify how causes contribute to an observed outcome are often termed *decomposition* or *integration functions*. Numerous studies have shown that intuitive causal judgments involving causes of binary outcomes conform to models that make the causal-invariance assumption better than those that do not (Cheng et al., in press; Griffiths & Tenenbaum, 2005; Park et al., 2022; see Lu et al., 2008 for a review).

However, reasoners frequently encounter causes that show a deviation from expectation assuming invariance. Causes may *interact* to produce an outcome (e.g., two medicines interact to cause a side effect). We report an experiment showing that, to reconcile aspiration with reality — "causal invariance as ideal" with "causal invariance as fallible default assumption" — reasoners apply the causal-invariance

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

assumption to interacting causes as a single unit, shifting the level of representation to a "*whole cause*" to preserve the aspiration toward useable causal knowledge.

The causal-invariance view

A basic tenet of cognitive science is that our perceptions and conceptions of reality are our representations, formulated in an infinite search space of possible representations (Fodor & Pylyshyn, 1981; Kant, 1781/1965; Kuhn, 1962/2012). The issue of under-determination pervades cognition (Atlas, 2005): in causal inference, multiple causal hypotheses can explain any observed pattern of events. Cheng and Lu (2017) proposed that searching in an infinite space with a necessarily limited set of candidate hypotheses requires a signal to know when to look outside one's current candidate set, to formulate or reformulate concepts and variables.

Because unknown or unobserved background causes of an outcome can occur, and may do so differently across contexts, this view posits that one such signal would be deviation from *causal invariance* --the unchanging operation of a causal mechanism -- when one applies causal knowledge. Because contexts may differ in background causes, a reasoner who applies causal knowledge inherently assumes as a default that a cause inferred from a learning context is invariant across the learning and application contexts.

If invariance is a default assumption of transfer, then reasoners must also have means of detecting deviation from invariance, as a signal to revise. This signal requires knowledge of *causal-invariance functions*, that is, knowledge of how causes of an outcome *would* combine their influences *if each causal mechanism operates unchanged in the presence of other causes*: One cannot know whether an observed outcome indicates a causal interaction unless one first knows *what* outcome would *mean* "no interaction". Such knowledge is *analytic* rather than *empirical* (cf. Hume's, 1739, "relation between ideas" and "matters of fact"; cf. Shepard's, 2008, distinction between logic and mathematics on one hand and empirical science on the other). Experience tells us how specific forces do combine their influences, but not how their influences *would* combine *if the causes operate invariantly*. Analytic knowledge is justified by reason: in the case of causal-invariance functions, by what logically follows from the "*sameness* of the causal influences" across contexts (for examples of work on probabilistic causal-invariance functions, see Cheng, 1997; Cheng et al., in press; Park et al., 2022). Empirical knowledge, by contrast, is justified by experience/observations, and transfer based on empirical knowledge is justified by similarity between the observations.

The causal invariance view also incorporates a part-whole distinction for the level of variable representation. The reasoner's aspiration is to formulate a *"whole" cause* variable that is assumed to apply invariantly from the learning to application context. Whole causes may be *elemental* or *complex*. Complex causes consisting of *interactive components*/parts (Novick & Cheng, 2004) are common. Playing a guitar chord, for example, requires a conjunction of finger placements and strumming. The components are individually insufficient but necessary parts of a collectively sufficient "whole" cause of the outcome (cf. Mackie's, 1974, *INUS* condition).

Whole causes—whether elemental or complex—are assumed to apply invariantly, until a deviation from the expected invariance prompts a need for knowledge revision. Revision involves changing the representation (see examples discussed in Cheng & Lu, 2017; Cheng et al., in press; Lien & Cheng, 2000; Park et al., 2022). A musician would typically play the F major chord on multiple guitars using the same configuration of finger placements for that chord (the interactive components represented as a complex whole that is denoted by a single variable, 'F major scale'), and assume that the configuration would produce the same set of musical notes across guitars. A flat note, however, a deviation from this default, provides a signal to revise the whole cause to include standard tuning as an additional component.

Generalization of integration functions: A test of two hypotheses

Previous research that studied integration functions has focused on the role of experience at two levels: fitting an adopted function to data and learning a function based on data (i.e., empirically). At the former level, the function is fixed and only weights can be learned based on data (e.g., the additive function in Rescorla & Wagner, 1972; the logistic function in logistic regression; the noisy-OR/causal-invariance function in Buehner et al., 2003). At the latter level, reasoners induce from data various *empirical integration functions* that describe how given causal variables combine to produce an outcome (e.g., Beckers et al., 2005; Lovibond et al., 2003; Melchers et al., 2004). Work at this level assumes fixed, pre-defined variables as input to the learning system, with the goal of learning an integration function that minimizes prediction error in terms of these variables, and notably does not include any analytical notion of causal invariance.

The causal-invariance view incorporates empirical learning but proposes a third, hierarchically higher level, at which analytic knowledge of causal-invariance functions is applied to whole causes. The empirical-function learning system works well to fit a function to the observed data — but without the analytic level, when faced with new combinations of stimuli, this system would have to base generalization on the acquired empirical function and similarity between test and familiar stimuli.

To test the empirical-function learning hypothesis against the causal-invariance hypothesis, we designed an experiment requiring reasoners to spontaneously distinguish between analytic and empirical knowledge of integration functions, with the goal of testing whether they apply the respective functions appropriately to their encoded whole causes and

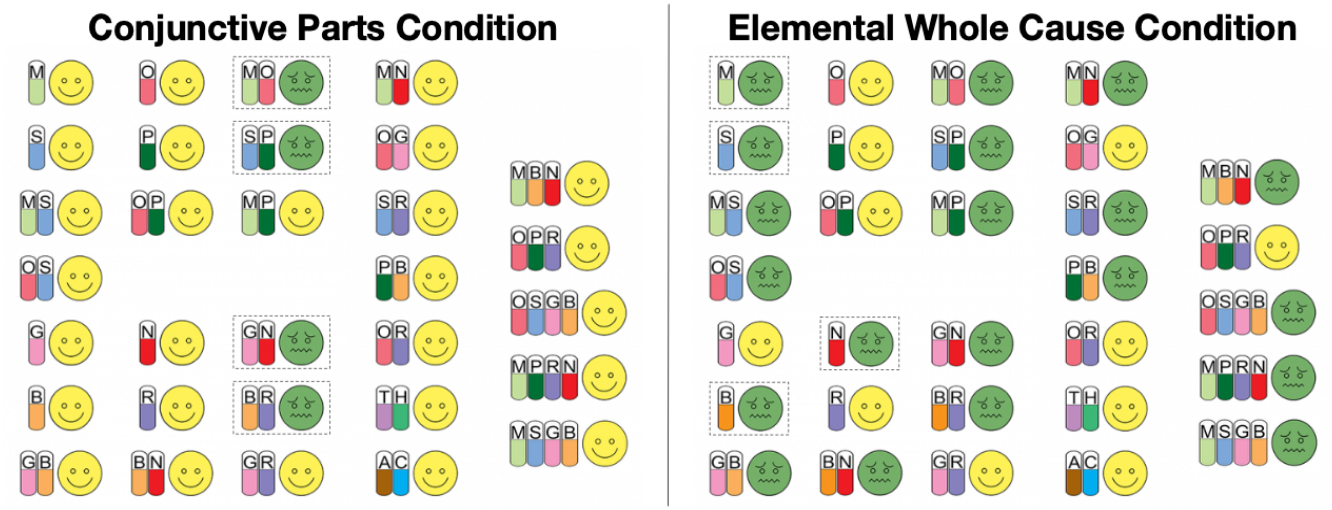

Figure 1. Summary of all 31 learning trials for each condition. The dashed boxes highlight the intended 'whole causes', which were not differentially labeled or indicated in the experiment itself.

interactive parts. Participants observed episodes in a fictitious scenario involving candidate causal stimuli, some of which were paired with the target binary outcome.

The experiment manipulated the integration function generating the outcome-frequency patterns: the scenario contained candidate causes that either 1) produce the outcome *only by interacting* with another candidate (the *Conjunctive* condition implementing a conjunctive-integration function) or 2) always *independently* produce the outcome, that is, are causally invariant across the presence or absence of other candidate causes (the *Elemental* condition implementing a disjunctive-integration function). (See Figure 1 for trials in the two conditions.) We measured transfer of the respective empirically learned integration functions to combinations of novel stimuli and to novel combinations of old stimuli, all within the familiar learning domain.

Both the empirical-function learning and causal-invariance hypotheses predict that participants will empirically learn whether causes obey conjunctive- or disjunctive-integration functions, depending on their condition's learning data. However, the two views differ in their predictions for novel combinations involving conjunctive causes.

Importantly, the previous literature embodying the empirical-function learning view (e.g., Beckers et al., 2005; Lovibond et al., 2003; Melchers et al., 2004) does not make explicit predictions for combinations involving conjunctive causes, as they do not test or address such a scenario. If predictions are based solely on the learned empirical functions, without causal invariance as an aspiration, then reasoners would continue to apply the empirically-learned function (*conjunction*) if they generalize that function; and if it does not, then it could either withhold judgment or resort to the base rate of the outcome. This ambiguity in prediction is implicit in the literature, and we return to this below.

By contrast, causal invariance predicts that a conjunctive cause would combine invariantly (*disjunctively*) with non-causes or other whole causes. In other words, the 'whole cause' representation, consisting of interacting parts, is assumed to obey invariance at the whole-cause level. Reasoners would thus apply the disjunctive-integration function to a conjunctive cause in combination with other whole causes and/or non-causes, even though in their experience *all* causes in that domain are conjunctive, not disjunctive. This transition away from the empirically-learned conjunction to *disjunction* highlights the role of analytic knowledge (justified by reason rather than experience).

Method

Participants

Participants were recruited through Prolific (in the US) and paid $8 USD for participation (approx. 40 mins). We estimated our targeted sample size in a preregistered power analysis (https://osf.io/35z4c) used for our previous study with similar measures (Bye et al., 2021). Of the participants based on our preregistered target, 254 passed cover-story comprehension checks and were randomly assigned to either the Conjunctive or Elemental condition.

Materials

The experiment was conducted using a Qualtrics online survey with a cover story, learning trials, attention checks, memory checks, and transfer questions.

Cover Story Participants were told to play the role of a medical researcher trying to determine whether or not their company's mineral supplements cause stomachaches as a side effect. The instructions clarified that the outcome was probabilistic, to appear realistic in order to engage intuitive reasoning: when a supplement intake (an individual supplement or a combination of supplements) causes a stomachache, then at least 85% of patients show the effect, but when it does not cause a stomachache, none of the patients do. After reading the description, participants were

given 5 true-or-false questions to ensure they comprehended the key elements. For example, a true item was: supplements were taken either individually or in combination. A false item stated that the survey is a game (included to reinforce that participants should not suspend their intuitive reasoning). Only participants who passed the comprehension check were allowed to continue.

Learning Stimuli and Attention Checks In both conditions, participants observed a total of 31 learning trials, one at a time, with each trial consisting of all patients taking either a single supplement or a combination of supplements, paired with the outcome. Stimuli were represented as a diagram with a corresponding verbal description of the trial, e.g., "Most patients got a stomachache after each intake of supplement M" (which we denote in-text as {M+}, where '+' indicates the presence of stomachache) and "The patients did not get a stomachache after any intake of supplement O" (denoted {O–} where '–' indicates the absence of stomachache). All diagrams contained pill-shaped supplement(s) with randomly-assigned colors and letter labels, accompanied by either a yellow "happy" face (no stomachache) or green "sick" face (stomachache). The supplement sets were identical across conditions (see all 31 trials in Figure 1).

To demonstrate the different intended integration functions for each condition, the identical supplement sets yielded different patterns of outcomes (Figure 1). Specifically, in the Elemental condition, taking two non-causal supplements together was never paired with a stomachache, and any supplement set that included at least one causal individual supplement would be paired with stomachache. (i.e., disjunction). The Conjunctive condition was identical to Elemental except that supplements combine their effects in an *interactive* manner to bring about the outcome: no individual candidate cause on its own was paired with the outcome, and only specific combinations of two supplements were paired with the outcome. For example, when taken individually, supplements M and O were not paired with stomachache {M–, O–}; taking them in combination, however, was paired with stomachache {MO+}. Crucially, the interactive combinations were *never* paired during learning with any other stimuli, so participants received no empirical evidence for how complex causes combine, leaving this entirely open to participants' own assumptions.

Both conditions were designed to have 4 "whole causes" (see stimuli enclosed by dashed boxes in Figure 1), which we define as the minimal (most parsimonious) set of supplement(s) that produce a stomachache. For example, in the Elemental condition, MO+ does not represent a whole cause because it contains O in addition to whole cause M. In Conjunctive, the 4 whole causes (e.g., MO+) were the only supplement combinations that on their own cause stomachaches. To rule out the inference that more supplements *per se* causes stomachaches, both conditions presented trials with 3 or 4 supplements and no stomachache (each condition's rightmost column in Figure 1).

Figure 2. Schematic of the "double decomposition transfer" trial with novel supplements J and K. Top: first, participants predicted whether most patients would get a stomachache after taking the combination of novel supplements J and K (JK?). Bottom: after their prediction for JK, they saw the novel results {J–, JK+} and predicted whether most patients would get a stomachache after taking K alone (K?).

To promote learning, interwoven with the learning stimuli in each condition were 11 attention checks asking whether a certain supplement set would cause stomachaches if taken again by the same patients. To promote generalization of the integration function, 4 more attention checks asked participants to predict the outcome for 4 *novel* combinations of two individual supplements they had only seen separately so far; they were then shown the real outcome for their condition. After completing all learning stimuli and attention checks, participants were given a memory check. In both conditions, the criterion was correctly recalling the outcome for the 4 true whole causes and 2 non-causes. All participants were allowed up to 8 attempts to pass the memory criterion.

Transfer: double decomposition Our primary transfer trials assessed both the learning and transfer of the respective empirical causal integration functions in each condition. (Additional transfer trials replicated items from Bye et al., 2021.) Participants were tested with new stimuli similar to those in the learning trials (introduced as "new but related mineral supplements") and asked to make predictions.

The trial (see Figure 2, top) began with a pair of novel stimuli (supplements J and K) presented in combination with a familiar supplement combination O and P. All participants were presented with new clinical results showing that most patients had a stomachache after intake of all four supplements {OPJK+}. They were also shown that, as would be expected based on the learning trials, no stomachache followed the OP combination alone {OP–}. To focus on inference by removing memory effects, both groups were shown a summary of their learning trials (as in Figure 1).

The first transfer question asked: based on the new clinical results {OPJK+, OP–}, would most patients from this study get a stomachache if they take supplements J and K together (JK?). As in Bye et al. (2021), participants responded first by answering either 'Yes' or 'No', and then indicated their confidence in their rating on a 5-point Likert-type item (see labels in Figure 3). Following preregistration, primary

analyses treated ratings as dichotomous Yes/No, with confidence ratings included for descriptive purposes.

After participants answered the first transfer question, they were then shown the results for that JK combination {JK+} as well as J alone {J–}. Analogously, they were then asked to predict the outcome for most patients from K by itself (K?), with the same response options. This question assessed whether the two groups did learn their respective functions.

These questions, similar to others' empirical-rule transfer questions (Melchers et al., 2004; Bye et al., 2021), both assessed how a reasoner decomposes an observed target outcome—the "total" outcome (e.g., {JK+} for the second transfer question) occurring in the presence of a known non-cause {J–} and an unknown candidate (K?)—into the inferred influences from constituent causes. In essence, the questions asked: what *explains* the occurrence of the outcome? The key difference between the questions was whether the unknown candidate was, for the Conjunctive group, at the level of a hypothesized whole-cause (supplement pair JK) or an interactive component (individual supplement K) according to the causal invariance view. The Elemental group served as a control. The double-decomposition trial thus controlled for all extraneous variables, including prior experience, uncertainty regarding the novel stimuli, and question format.

For the {J–, JK+, K?} item, both the empirical-function learning view and the invariance view predict transfer of each group's respective learned integration rule to the novel stimuli. They thus both predict opposite responses for the Conjunctive and Elemental groups. Consider participants who learned their group's integration rule and are willing to generalize it to novel items. For these Conjunctive participants, supplement K would be merely an interactive component that does not produce stomachache on its own. Both views therefore predict K– for Conjucntive. In contrast, for Elemental participants, supplement K by itself must be a whole cause. Thus, both views predict K+ for Elemental.

The views differ, however, in their predictions for {OP–, OPJK+, JK?}. Consider the causal view. Because this view predicts that causal invariance would be the default decomposition function at the whole-cause level regardless of the empirically acquired integration function, it predicts that both the Elemental and Conjunctive groups would respond that the patients would get a stomachache after taking JK by itself, despite their different empirical inferences about what constitutes a whole cause (individual supplements vs combinations of two supplements). More specifically, in order for the Conjunctive group to explain {OP–, OPJK+}, decomposition assuming invariance (disjunction) predicts that the JK combination *must be* a whole cause of stomachaches. Otherwise, there would be no whole causes at all in the OPJK combination to explain the occurrence of the outcome. The Elemental group would likewise use invariance as their decomposition function: based on their experience, they would infer that at least one of J or K (individually) is a whole cause of stomachache.

By contrast, while the empirical-function learning view (like the causal-invariance view) predicts that Elemental participants would predict JK+ for the first transfer question (since JK must contain at least one elemental whole cause, namely, J, K, or both), this view predicts that Conjunctive participants would either generalize their conjunction rule (in which case {OPJK+} decomposes into conjuncts {OP–} and {JK–}), withhold judgment, or resort to the base rate of the outcome (most supplement pairs are non-causal). Thus, lacking the use of analytic causal invariance at the whole-cause level, this view predicts that Conjunctive participants will answer JK– (by either conjunctive decomposition or base rate) or not at all. Put another way: no purely empirically-based view of causal induction would predict that the Conjunctive group would apply different decomposition functions in reply to the two transfer questions. Both questions concern decomposition involving the same two novel supplements J and K, from the same domain, with identical relevant prior causal knowledge. It is the causal invariance view that explains how and why the Conjunctive group would apply disjunction at the whole-cause level (JK+) and conjunction at the interacting-parts level (K–).

Procedure

Participants were recruited through Prolific and forwarded to a Qualtrics survey. Only participants who correctly answered cover-story comprehension checks within 2 attempts continued to be randomly assigned to a condition. Only those who passed memory checks received final transfer questions. Participant recruitment was stopped after 254 participants had completed the study.

Results

All statistical analyses were conducted in R and followed the above preregistration, except as clearly noted below. All transfer items from Bye et al. (2021) replicated the earlier results, so for space we focus here on the novel items.

For all 254 participants who passed the inclusion criteria, we used our preregistered criteria for categorizing participants as having inferred the intended integration function (disjunction or conjunction) in two ways, one stricter than the other. The 3 criteria consisted of 1) transferring the intended empirical function to novel items within the domain, and 2) agreeing with probe questions consistent with the intended empirical function, but 3) disagreeing with those consistent with alternative functions. Out of 126 Elemental participants, 108 satisfied all 3 criteria, and out of 128 Conjunctive participants, only 18 satisfied all 3 criteria (indicative of the complexity of conjunction). Due to the relatively small subset of Conjunctive participants who passed all 3 criteria, we analyzed the data in two subsets: participants who passed All Criteria (as preregistered) and the larger subset who at least passed the most basic criterion of transfer. Note this transfer criterion, while less strict, is analogous to the criteria from previous studies (e.g., Lucas et al., 2014; Melchers et al., 2004).

We analyzed the double-decomposition trials among the

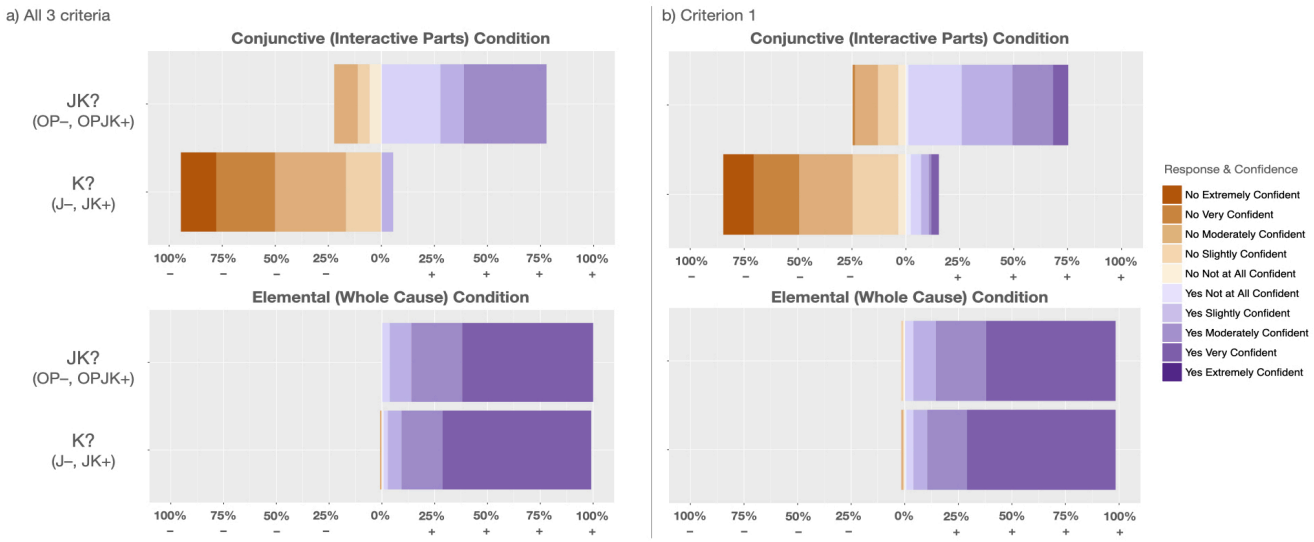

Figure 3. Stacked Likert plots for percent of participants' responses in each condition to the double-decomposition transfer items (Figure 2). The width of each colored bar represents the percentage of participants who answered Yes (purple) or No (orange), with the color intensity representing their confidence rating. The more purple (and the more the bars are shifted to the right), the higher the proportion predicting a stomachache, as represented by the x-axis labels.

All Criteria and Criterion 1 subsets (Figure 3). For each condition, the proportion of participants predicting a stomachache for supplement combination JK after observing {OP–, OPJK+} are shown in the top bar for the condition. Consonant with the "whole cause" prediction according to the causal-invariance view, participants meeting All Criteria (left side of figure) reliably predicted a stomachache for JK in *both* conditions (top and bottom of figure), with 77.8% of Conjunctive participants (binomial $p = .031$) and 100% of Elemental participants ($p < .001$). Similar results were found for the larger subset passing at least Criterion 1 (both binomial p's $< .001$).

The proportion of participants predicting a stomachache for K alone after being told that JK did lead to a stomachache but J alone did not {JK+, J–} are shown in the bottom bar for each condition in Figure 3. Here, strongly in support of the transfer of each group's acquired integration function, participants qualitatively differed by condition: among those meeting All Criteria, 99.1% of Elemental participants predicted a stomachache for K alone ($p < .001$), while only 5.6% of Conjunctive participants did ($p < .001$), which was a significant difference between conditions (Fisher's exact test, $p < .001$, $OR = 1132.62$). (A similar difference between conditions was found for the Criterion 1 subset, $p < .001$, $OR = 317.63$.)

For the Conjunctive condition, participants were significantly more likely to predict a stomachache for JK than for K alone (McNemar's p's $< .001$ for both subsets). By contrast, for the Elemental condition, there was no difference between participants' predictions for JK and K (McNemar's ps $= 1$ for both subsets). The switching of decomposition functions from disjunction at the whole-cause level (JK) to conjunction for interactive components (K) in the Conjunctive group but not the Elemental group is just as predicted by the causal invariance view, but not explained by any purely empirically-based view.

Discussion

Our results provide support for the causal invariance view, suggesting that the goal of constructing useable causal knowledge constrains human causal induction. Participants generalized empirical integration functions to novel stimuli in the same domain, yet they spontaneously defaulted to using *causal invariance* functions at their "whole cause" level, supporting a role for *analytic knowledge* as a guide in causal learning over and above empirical learning. Specifically, the Conjunctive group generalized the empirical conjunctive rule involving interactive components to decompose a novel supplement pair at the component (individual supplement) level—yet they applied *invariance* (i.e., disjunction) to decompose a novel four-supplement combination at the whole cause (supplement pair) level. This pattern of response is inexplicable by merely applying an empirical function with no distinction between the two levels (empirical-only view). Uncertainty due to lack of experience cannot explain why Conjunctive participants switched from generalizing their empirical function to applying invariance for the same supplement set featuring novel items. Our results corroborate our previous findings (Bye et al., 2021).

In our view, the causal induction process that navigates the vast search space of representations is guided by analytic knowledge of causal invariance functions. This process embodies a rational solution to the causal-induction problem that humans face: in order that we may know how nature works, our induction process assumes that the causal world is composed of invariant causes, and its aspiration is to construct representations of such causes.

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: LMC, 31*, 238-249.

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.

Bye, J. K., Chuang, P.-J., & Cheng, P. W. (2021, July). *When and why do reasoners generalize causal integration functions? Causal invariance as generalizable causal knowledge.* Poster presented virtually at the 43rd Annual Meeting of the Cognitive Science Society. Conference abstract available (p. 3236): *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological review*, *104*(2), 367.

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., Sandhofer, C.M., & Liljeholm, M. (in press). Analytic causal knowledge for constructing useable empirical causal knowledge: Two experiments on preschoolers. *Cognitive Science.*

Fodor, J. A. & Pylyshyn, Z.W. (1981). How direct is visual perception? Some reflections on Gibson's "ecological approach". *Cognition*, *9*, 139-196. doi: 10.1016/0010-0277(81)90009-3

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51*(4), 285-386.

Hume, D. (1739/1987) *A treatise of human nature* (2nd edition, Clarendon Press, Oxford).

Kant, I. (1781/1965). *Critique of pure reason*.

Kuhn, T.S. (1962/2012). The structure of scientific revolutions (4th edition). Chicago: U of Chicago Press.

Lien, Y., & Cheng, P.W. (2000). Distinguishing genuine from spurious causes: A coherence hypothesis. *Cognitive Psychology*, *40*, 87-137.

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

Lu, H., Yuille, A. 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. L., Gopnik, A. (2014). When children are better (or at least more open-minded) learners than adults: Developmental differences in learning the forms of causal relationships. *Cognition*, *131*, 284-299.

Mackie, J. L. (1974). *The cement of the universe: A study of causation*. Oxford, England: Clarendon Press.

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.

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

Park, J., McGillivray, S., Bye, J. K., & Cheng, P. W. (2022). Causal invariance as a tacit aspiration: Analytic knowledge of invariance functions. *Cognitive Psychology*, *132*, 101432.

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.

Shepard, R. (2008). The step to rationality: The efficacy of thought experiments in science, ethics, and free will. *Cognitive Science*, *32*, 3-35.

van Rooij, I., & Baggio, G. (2021). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. *Perspectives on Psychological Science, 16*(4), 682-697.

Woodward, J. (2000). Explanation and invariance in the special sciences. *British Journal of the Philosophy of Science*, *51*, 197-254.

Woodward, J. (2010). Causation in biology: stability, specificity, and the choice of levels of explanations. *Biological Philosophy*, *25*, 287–318. DOI 10.1007/s10539-010-9200-z