Revisiting Hume in the 21st century: The possibility of generalizable causal beliefs given inherently unobservable causal relations

Nicholas Ichien and Patricia W. Cheng

University of California, Los Angeles

Corresponding author: Patricia W. Cheng

Email address: cheng@lifesci.ucla.edu

Invited chapter in A. Wiegmann & P. Willemsen (Eds.), *Advances in Experimental Philosophy of Causation*. London, UK: Bloomsbury Press.

Abstract

This chapter revisits two issues raised by the philosopher David Hume: 1) causal relations are beliefs formed in the mind based on inherently noncausal data (Hume, 1739/1987), and 2) experience is useful only if the future resembles the past (Hume, 1748/1975). These issues respectively specify the input and output for a rational causal-induction process. Together they raise the question: How is it possible to tease apart a target candidate cause's influence from that due to background causes, in a way that yields causal knowledge that generalizes across the learning and application contexts? The first half of the chapter reviews Hume's first issue from multiple contemporary perspectives – cognitive psychology, cognitive neuroscience, perceived causality in different contexts and inertial reference frames, and cognitive causal "illusions". We explain why, in view of the noncausal input and the desired output – useable causal beliefs – the causal-belief formation and revision process requires the *causal invariance* constraint: an assumption that there are causes that operate the same way in an application context as in the learning context. Causal invariance, along with parsimony, constrain the search for useable causal beliefs in the vast empirical space of possible representations. The second half of the chapter presents an argument showing that, without the causal-invariance constraint, intuitive causal induction and normative statistical inference would both fail to aim at generalizable causal beliefs. To our knowledge, Hume's two issues have not been examined from the perspective of cognitive constraints that would be conducive to arriving at generalizable causal knowledge.

Keywords: causal induction, decomposition function, invariance, generalizability, statistical inference, observability of causal relations

The present chapter is an introduction to a basic problem in causal induction: how is generalizable causal knowledge possible? We first present the problem of causal induction as posed by David Hume (1739/1987, 1748/1975) and clarify three confusions surrounding this problem. We go on to review empirical evidence from four perspectives that all provide support for Hume's view: Causal relations are not in the input to the reasoner/cognitive system. What is in the input are merely the states of the candidate causes and the state of the outcome-in-question due to all its causes present in the context. Making a causal inference about a target candidate cause therefore requires decomposing the observed outcome into contributions from the target cause and from other causes of the outcome in the context. Nature does not tell the reasoner how to decompose the observed outcome — the task is up to the reasoner. After establishing the problem of causal induction from the perspective of cognitive science, we explain what the assumption of causal invariance is and why it is a necessary constraint for rational causal induction. We end the chapter by relating our analysis of causal invariance to normative causal inference in statistics. Our chapter does not assume any background knowledge of work on causal induction in philosophy or psychology. Our intention is for it to be of interest to upper-level undergraduate students, graduate students, and anyone else who enjoys thinking through the problem our mind solves when it aims to infer a generalizable causal relation.

People often have the compelling intuition that they directly "see" causation, and thus have no need to *infer* causation. If they see an unfortunate person killed by a volcanic eruption, overtaken by a pyroclastic flow, it may seem hard to deny that they perceived the reality of the volcanic eruption killing the person.[1] If they see a moving ball hit a stationary ball, and the stationary ball starts to move away, they "see" the true "launching" into motion of one ball by motion in the other. If their right-hand fingers scratch a mosquito bite on their left arm, and their left arm feels relief from the itch, they directly perceive the relieving of the itch by their scratching.

Hume (1739/1987) argues that counter to our compelling intuition that a moving ball launches a stationary ball when we observe the former hit the latter, the causal aspect of that intuition is an inference in our mind and is absent in the observation itself. Hume (1748/1975, p. 37) also brings attention to an assumption so intuitive that we may be unaware of making it:

[1] Our thanks to an anonymous reviewer for the example.

Whenever we generalize from a learning context to an application context, we assume, "the future will resemble the past." He goes on to state its implication, "If there is any suspicion that the course of nature may change, … all experience becomes useless …" Together, Hume's two points raise the question: if causal perceptions and beliefs are mental constructs absent in the observations in our experience, on what basis would one expect these mental constructs to capture the unchanging course of nature, such that experience is not useless?

Three confusions clouding the nature of the problem of causal induction

It is tempting to conclude that the compelling perception of causal relations renders inference unnecessary, at least in cases in which causation appears "observable". But the conclusion that causation is observable involves three sources of confusion.

The first is a confusion between the input of the causal induction process and its output. The conclusion mistakes the compelling perception of causation, as illustrated in our examples, to be the *input* to the causal induction process, when it is in fact the *output to be explained* (see Henle, 1962, for an example in which confusion around the input to a cognitive process, deductive inference in her case, creates confusion about the process itself). This confusion may be due to the vagueness of Hume's criterion for what he does or does not find "evident" in the observations (1739/1987, pp. 649-650). In contemporary information-processing language, a paraphrase of Hume's thesis that causal relations are not evident in the observations would be: Causal relations are not in the input available to a process that infers cause and effect – the construct 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 arriving at the causal output from its noncausal input.

The unobservability of causation is a specific form of the general challenge of formulating adaptive knowledge: reality in the world does not come represented (Goodman, 1955; Hawking & Mlodinow, 2010; Kant, 1781/1965). All our perceptions and conceptions of reality are our representations of it, formulated within an infinite search space. Consider our perception of a cube. The 2-dimensional image cast by a cube on our retina is ambiguous in that it can map onto an infinite number of differently shaped 3-dimensional objects (e.g., see Pizlo, 2001). Yet, despite the inherent under-determination of the distal object, we 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).

Thus, with respect to the stereoscopic vision process, 3-dimensionality is "unobservable"— a shorthand for "being absent in the input to a process" — and is the to-be-explained output of the process. Likewise, causation is "unobservable" with respect to the causal induction process, and the perceived "necessary connection" between a cause and an effect is the to-be-explained output (Hume, 1739/1987).

A second source of confusion is that the apparent examples of observable causation often involve prior causal knowledge at a more abstract level than the particular causal relation in question. Although a reasoner may be witnessing a pyroclastic flow hitting someone for the first time, they almost certainly know, at a more general level, from knowledge of landslides and fires, that being struck by massive flows of hot or heavy matter can be fatal. Lien and Cheng's (2000) hierarchical consistency hypothesis explains how consistency and inconsistency of covariations between potential cause and effect variables across representations at different levels of abstraction can explain conclusions of causality or noncausality ostensibly based on a single instance. Their paper presents evidence showing that information beyond what is in the single instance gets recruited, more specifically, that judgments involving a single instance can be explained by retrieval from causal schemas in long-term memory formed by past causal inferences, rather than by "observable" causality. [See Rips (2011) for a review of evidence and arguments against perception of causality as the source of the causal knowledge.]

A third source of confusion is that examples of observable causation concern situations in which only one cause is perceived to be present (i.e., the reasoner assumes no background causes). In such cases, an inferential process, either inductive (Cheng, 1997) or deductive, can account for the causal percept. Deduction such as the following would reach our intuitive causal conclusions:

Premises: 1) effect *e* occurred in situation *x*

2) effects do not occur without a cause

3) *c* is the only candidate cause in situation *x*

Conclusion: *c* caused *e*. In other words, the fact that we humans are able to judge causation in situations involving one single plausible cause does not imply that we do not have a

general causal-induction process capable of inferring new causal knowledge in situations involving more than one plausible cause. The single-cause situation may be regarded as a trivial case of the application of that inference process.

To illustrate that causation in situations with a single plausible cause is not observed but inferred, we review the striking "phantom hand" phenomenon (Armel & Ramachandran, 2003; Botvinick & Cohen, 1998; Ramachandran & Hirstein, 1998). More generally, the phenomenon is a good reminder of the inferential nature of our conception of reality. Armel and Ramachandran report that participants with normal sensation and perception perceived touch sensations as arising from a rubber hand. This occurred when both the rubber hand in view and participants' own out-of-view real hand were repeatedly tapped and stroked in a random sequence in synchrony. In other words, participants perceived the tapping and stroking of the rubber hand as causes of their perceived touch sensations, as if the rubber hand is part of their body. An analogous illusion was obtained even when a table top was similarly tapped. The perceived causal relation could not have been an "observation" of causation, because no such actual causation existed in the experimental setup. The perception was so internalized that when the rubber hand or table was then "threatened" with potential injury, participants winced and sweated. It was as if the participant perceived a threat to the table as a threat to their hand. Consistent with their behavioral response, participants displayed a strong skin conductance response (SCR)[2] in the real hand, even though no threat was issued to it. Notably, when there is only one plausible cause of our sensations, even something so fundamental as the perceived boundary of our body is mutable to allow attribution of the sensations to that single cause. Armel and Ramachandran write (p. 1499), "one's body image is itself a 'phantom': one that the brain constructs for utility and convenience".

Thus, for the hypothesis of "observable causation" to be tenable, processes such as deduction and the recruitment of prior causal knowledge must be ruled out as explanations of the causal conclusion. We propose that, to provide clear evidence for observable causation, the critical discriminating test is to compare causal judgment between two situations: A) a single-cause situation and B) a situation in which a second cause is introduced without disturbing the causal sequence in situation A. When multiple plausible causes are present, from either current

[2] SCR is a physiological measure of psychological and autonomic arousal that is not under voluntary control.

information or prior knowledge, it would no longer be possible for deduction to narrow down to one cause as the compelling conclusion. But if a causal relation is "observable", the relation should be just as discernable, whether there is one cause or two causes present. If we can see an apple in a bowl, we should still be able to see it when another apple is placed in the bowl. We presently review some empirical evidence comparing the two types of situations.

Four perspectives in support of the unobservability of causal relations (Hume, 1739/1987)

Now that we have clarified the three common sources of confusion, let us turn to empirical evidence in favor of causal relations being unobservable rather than observable. We examine this issue from four perspectives: 1) psychological evidence on the perception of causality when there is more than one cause, 2) the relativity principle with respect to the invariance of laws of motion across inertial reference frames (Newton, 1687/1713/1726/1999), 3) the visual input to our cognitive system and its relation to the cognitive neuroscience of color perception and, by extension, causal perception, and 4) causal inference about internal psychological outcomes.

Perspective 1: When a second cause is introduced, the compelling perception of causation disappears.

We compare two situations below, a single-cause and a two-cause situation, summarizing the discussion of them in Cheng (1993).

Michotte's (1946/1963) Experiment 21 provides a clear demonstration that perceived launching is not the result of a direct perception of causation. Because Experiment 21 differs from the basic version of Michotte's often-cited launching experiments in only one respect, we first describe the basic version in Experiment 1. In Experiment 1 (see top sequence in Figure 1), an Object B is in the middle of the screen, and the subject fixates on it. At a given moment, Object A enters the screen from the left and moves toward B at a constant speed. Object A stops at the moment it comes into contact with B, while B then starts and moves to the right. The blow by A is often perceived to "send B off," "transfer its momentum to B," "cause B to move," or "launch B".

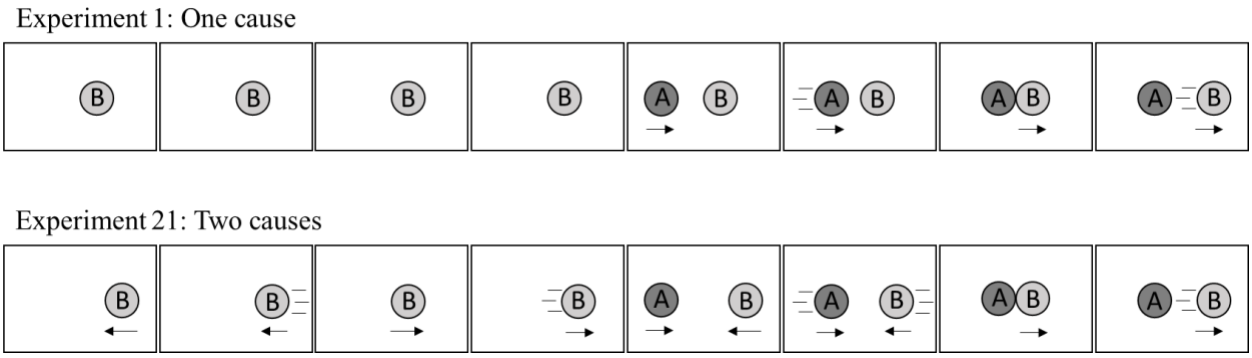

Figure 1. Illustration of Michotte's (1946/1963) Experiments 1 (top sequence) and 21 (bottom sequence). The visual stimuli are identical in the two experiments from the moment of impact on (represented in the two rightmost frames in the top and bottom panels). Arrows under Objects A and B indicate motion in the direction of the arrow (see a video of the two demonstrations at https://youtu.be/ZVZpggGXl08).

Describing his results when the collision satisfies some fairly strict spatio-temporal constraints (e.g., B has to start moving within a few tenths of a second of the collision, A cannot stop before touching B), Michotte (1946/1963) writes, "The impression is clear: it is the blow given by A which makes B go, which produces B's movement" (p. 20). He regards his findings as refuting Hume's (1748/1975, Section VII, part ii, p. 74) claim that "we never can observe any tye between 'a cause and an effect'". Instead, the cause that produces motion can be "directly experienced" (Michotte, p. 21).

In the launching phenomenon, the effect may be characterized as "Object B moving away starting at the moment of collision," and the cause may be characterized as "Moving Object A hitting Object B." Note that the input to the perceptual process consists of the activities of a mosaic of photoreceptors that change over time, nothing more. "Launching" is not in the activity of any of the photoreceptors, and there is no homunculus. Experiment 21 shows results consistent with this information-processing perspective.

Experiment 21 illustrates the two-cause situation (see bottom sequence in Figure 1). One often overlooked finding is that the perception of launching is critically dependent on the state of B *before* A collides with it. The stimuli in Experiment 21 (Michotte, 1946/1963) is identical to those in Experiment 1, except that B moves to and fro before A enters. B's oscillation is timed so that A collides with B just as B comes to a rest and is about to move to the right. The sequence at impact and thereafter is identical to that in Experiment 1. Because the effect of impact by A on B

cannot precede the impact itself, B's to-and-fro motion prior to the impact cannot be part of the "effect" in question. Instead, it indicates a cause other than the blow by A.

Michotte (1946/1963) reports that there is no perceived launching with the set-up in Experiment 21: B's movements "seemed entirely independent of the movement performed by A" (p. 74). In comparison with Experiment 1, it is evident that B's to-and-fro motion prior to the collision eliminates the impression of launching. In view of the fact that the cause-and-effect sequences in the two experiments are identical, if causation is indeed observable, launching should be perceived equally in both experiments. The fact that it is not indicates that, even in the case of compelling causal perception, causality cannot be present in the sensory input. To conclude from the perception of causality in the launching phenomenon that causality is "observable" is to mistake the output of the perceptual system for its input. Crucially, that causal perception is truly a perceptual phenomenon is entirely consistent with our claim that causation is not observable, and we fully acknowledge that researchers of causal perception do not take the phenomenon as evidence for the observability of causation (e.g., Scholl & Tremoulet, 2000). Our present point is that this phenomenon necessitates explanation. Specifically, how is it that, in the case of causal perception, our perceptual system takes non-causal sensory input and reliably generates the compelling visual representation of "launching" in cases like Experiment 1 but not in cases like Experiment 21?

Perspective 2: If causation is not a mental construct, perceived causality should not change across inertial reference frames (Cheng & Lu, 2017)

We next consider the compelling perception of causality in ball collision episodes (Michotte's, 1946/1963) from the perspective of the postulate of relativity in Newtonian physics. Consider the perception of causality in each of three horizontal motion episodes involving the collision of two balls. Assume an idealized world in which there is no friction and no background scene to convey the position of the balls relative to the background. The issue concerns which ball is perceived as the cause of what happens in each of the episodes in Figure 2:

Figure 2. Three views of the same collision event from three different inertial reference frames (see a video of the episodes at https://www.youtube.com/watch?v=H7ukG3OAT7I).

Episode 1: Ball B is stationary at the center of the screen. Ball A appears from the left, moves toward Ball B with constant velocity *v* and collides with it. Ball A stops and B moves to the right with velocity *v*.

Episode 2: Now, Ball A instead is stationary at the center. Ball B appears from the right, moves toward A with velocity –*v* and collides with it. (The negative sign indicates movement from right to left.) B stops as A moves to the left with velocity *v*.

Episode 3: Balls A and B simultaneously enter from the left and from the right, respectively, *at half the speed* (*v*/2) as in the other two episodes. They collide, and move away in opposite directions at the same speed after their collision as before.

In accordance with Michotte's (1946/1963) findings, virtually everyone perceives that in Episode 1 Ball A "causes" Ball B to move. The reverse holds in Episode 2: here Ball B "causes" Ball A to move. In Episode 3, the perception is that each ball causes the other to rebound after their collision. If the balls were real objects rather than cartoons, the preceding perceptions of causality would hold just the same.

Although we perceive the three collision episodes as involving different configurations of causal roles, these episodes can depict the exact same event viewed from different inertial reference frames. An inertial reference frame is a system of coordinates that moves at a constant

velocity. A postulate in Newtonian physics is that laws of motion are invariant across inertial reference frames (Newton, 1687/1713/1726/1999).

To see the three episodes as views of an identical physical event from three inertial reference frames, imagine watching the top episode from clouds moving respectively with constant velocity *v* and *v/2*, one cloud at a time. The two clouds represent different inertial reference frames. Watching the top episode from each of these two "clouds" transforms that episode respectively into the middle and bottom episode. The exact same event necessarily involves the same causation. Shifting the viewpoint across three inertial frames does not change the event, because the laws of motion are invariant across such frames. But the two balls' causal roles are perceived to differ across episodes. If causation is observable, why would an identical event, involving identical causation, give rise to three compellingly different causal perceptions?

Our three episodes illustrate that, counterintuitively, even in this compelling case of colliding balls, our perception of causation is not a direct reflection of nature. Nature does not come defined by variables or concepts. The concept of an inertial reference frame, for example, is a human construct. Perceived or conceived causation is a matter of how our cognitive processes "choose" to represent reality, in everyday thinking and in science. Whereas intuitive constructs describe these episodes as different events involving different causal roles, Newtonian constructs treat the three episodes as equivalent. Newton's choice yields greater causal invariance, covering a broader explanatory scope (Woodward, 2000). Our example illustrates that the reasoner's goal cannot be to "accurately" represent reality. It is instead to construct more useful, more predictive representations of reality, so that experience is not useless.

Perspective 3: If an activated cone does not know which combination of photons activated it, can "launching" be present in the sensory input to our visual system? (e.g., Hofer, Singer, and Williams, 2005; Mitchell & Rushton, 1971)

From a cognitive neuroscience perspective, causal relations cannot possibly be in the input to our cognitive system. Consider the nature of the input to our receptors, the ultimate and sole source of information about the material world (ourselves included in the material world). Here we review findings on human vision, because sensory input to the visual system is precisely specifiable. The perception of color is perhaps even more compelling than that of causal relations. But the confusion between the input and the output of a system may be more tempting

for causal induction, as it may be easier to see that color is in our head, not in the electromagnetic waves.

It is common knowledge that daytime vision in normal human vision is based on the activation of three types of cones, photoreceptors sensitive to electromagnetic waves in the light spectrum. We denote them S, M, and L cones to indicate their respective maximum sensitivity to short, medium, and long wavelengths of light.

Each cone type is sensitive to a range of wavelengths, with overlap between their distributions of relative sensitivities (see Figure 3). The overlap in relative sensitivity between the M and L cones is especially large. For example, for rays of 550-nm light, M and L cones are both likely to be activated. Thus, although the cone types are activated with different probabilities by light of different wavelengths, the overlaps imply that when a cone is activated, it would not "know" which wavelength of light activated it.

As vision researchers Mitchell and Rushton (1971, p. 1041) note in their "*Principle of Univariance*", an activated cone *does not know* which combinations of photons activated it – *it only "knows" that it is activated and the intensity of its activation*. The distal stimulus is under-determined by the proximal stimulus: An infinite set of wavelength-intensity combinations of electromagnetic waves can elicit an identical response from a cone or a single type of cone. Stepping back from color perception to causal perception, in view of what an activated cone can only know, and thus what it *cannot* know, it should be clear that none of the activated cones can know that some object is "launching" another object.

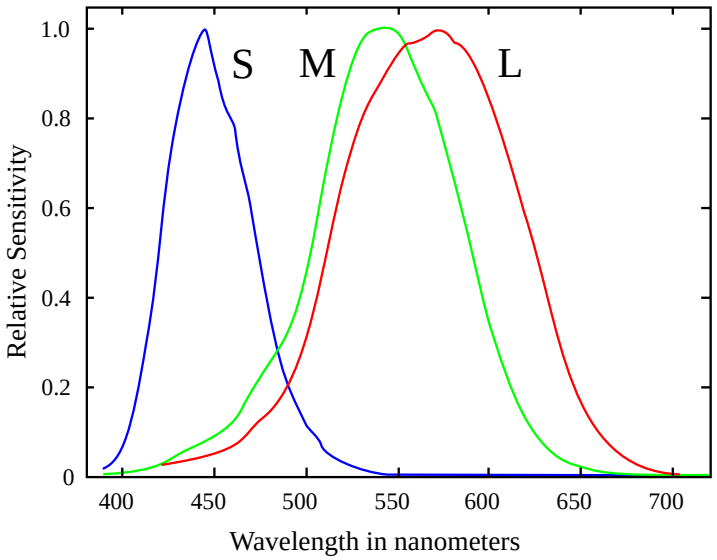

Figure 3. Distributions of the relative sensitivity of S, M, and L cones to electromagnetic waves with wavelengths in the range that give rise to human color vision. Peak sensitivities are normalized to 1.0.

Findings reported by Hofer et al. (2005) illustrate the remarkable vagueness of the color information encoded in each cone. When human subjects viewed a minuscule spot of 550-nm light that activates a single cone, so that S cones are unlikely to be involved, each subject gave a wide range of verbal responses across trials indicating their perception of color for the same 550-nm light (see Figure 4 from Hofer et al. below). The most common overall response is "white", in addition to at least 5 other color categories for each subject. Even though S cones are very unlikely to be involved in the detection, "blue" was a quite frequent response for 3 of the 5 subjects. If color is not represented in any cone, even less so are other features of our conception of the world; features such as causation, object-hood, and 3-dimensionality are not represented in any of the photoreceptors that inform our daytime vision. There is no homunculus downstream, only more neurons communicating with each other via synapses. Thus, from what is known about the nature of the sensory input, causality is not in the input to our cognitive system from the external world.

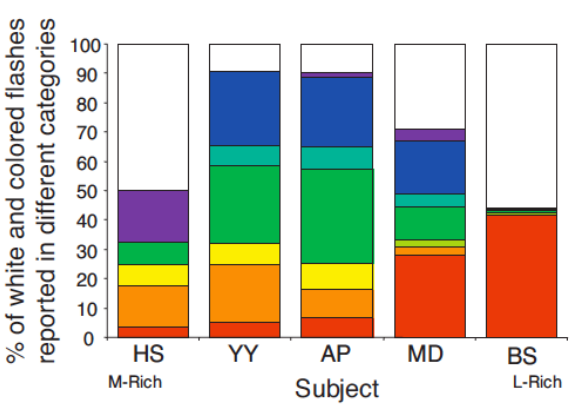

Figure 4. The color sensations reported by subjects when viewing a small spot of 550-nm light. At this wavelength only L and M cones participate in detection. Shown are the percentages of white and colored responses that were placed in each response category, interpolated at 50% frequency of seeing. From Hofer et al. 2005.

Perspective 4: People are unaware of the causes of internal psychological or physical outcomes when multiple plausible causes were present (e.g., Brasil-Neto et al., 1992; Nisbett & Wilson, 1977)

Inducing the causes of internal events is not different from inducing those of external events. Just as causation in external events is inherently not in the input to the processes that give rise to the causal understandings of those events, neither is causation in internal events in the input to the analogous processes. Recall that scratching a mosquito-bite involves proprioceptive input from the fingers, hand, and arm, together with visual input on the cones, enabling the integration across the inputs. No sensory receptor involved in the judgment "knows" the causal perception that the scratching relieved the itch.

If causation in internal events *were* present in the input, one would expect people to be aware of the causes that bring about their voluntary actions, not only in situations when there is one single plausible cause of an action present, but also when multiple plausible causes are present. To provide experimental evidence that causation in internal events is not in the input, we review some findings concerning multiple-cause type situations. In this critical test case, people are remarkably clueless about the causes of their voluntary actions.

In an experiment on the effect of transcranial magnetic stimulation on motor response, normal adult participants were asked to extend the index finger on either their left or right hand

at will (Brasil-Neto et al., 1992). They were asked to choose which hand to move upon hearing a go-signal. When magnetic stimulation was delivered to the motor area, participants more often moved the hand contralateral to the site stimulated. This response bias was independent of handedness and of the cerebral hemisphere stimulated. The researchers note that although the influence of magnetic stimulation on hand choice was clear and predictable, no participant was aware of the influence. They conclude (p. 964), "It is possible to influence endogenous processes of movement preparation externally without disrupting the conscious perception of volition." From their finding, we see that when there were two plausible causes of finger movement – magnetic stimulation and participants' own choice of hand – participants were unaware of the actual cause of their "willed" finger movement.

Similarly, in Nisbett and Wilson's (1977) classic article, "Telling more than we can know", they review numerous striking findings showing that people not only cannot articulate the causes of their behaviors and actions, but even when the true cause is revealed, people refuse to believe that such can be the case. For example, in a study conducted in a commercial establishment under the guise of a consumer survey, passersby were invited to appraise articles of clothing and choose one. In one condition, subjects saw four identical pairs of nylon stockings in an array and were asked to evaluate their quality. Once they announced a choice, they were asked to explain why they had chosen what they chose. There was a pronounced position effect, such that the rightmost item in the array was heavily over chosen. The right-most stockings were chosen almost four times as often as the left-most. When asked about the reasons for their choice, no subject ever mentioned the position of the stocking in the array. Even when asked directly whether they chose the article because of its right-most position in the array, "virtually all subjects denied it, usually with a worried glance at the interviewer suggesting that they felt either that they had misunderstood the question or were dealing with a madman" (p. 244).

These examples illustrate that people can be unaware of the causes of their decisions, in simple choices in everyday life or in the laboratory. To our knowledge, there has not been evidence showing that when multiple plausible causes were present, the internal causal relation was "observable".

The four diverse perspectives just reviewed converge in showing that causal understandings are our representations of reality, rather than "direct reflections" of reality. If causal understandings are not directly given by reality, then how and why do we humans develop

the causal representations that we do? We do so because we need generalizable/useable causal knowledge, and our causal-induction process aims at formulating such knowledge. The representational nature of causal knowledge implies that the search for such knowledge occurs in an infinite space of possible causal representations (recall the analogous issue in the perception of a cube). In the following, we address two questions arising from that challenge: 1) how is it possible to reduce the search space to avoid paralysis, and 2) how is it possible to tease apart a target candidate cause's influence from that due to potentially unobserved background causes? The rest of our chapter examines an answer to these questions in terms of a constraint on causal induction, which we term causal invariance (Cheng & Lu, 2017; Woodward, 2000).

Causal invariance as a rational constraint on causal induction

In the following, we explain how analytic knowledge of causal invariance plays an essential role in inducing *useable* causal representations, "analytic" in the sense that the knowledge logically follows from the meaning of the concept, namely, the *sameness* of causal influence across contexts, and "useable" in the sense that the acquired knowledge holds when it is applied. We do so by comparing causal induction that is constrained by causal invariance with an associative foil of causal induction that is not so constrained. In an extended example, we show that the former yields useable causal representations and that the latter does not. Because this difference occurs for causal representations involving discrete outcomes, and is simple to show for binary outcomes (e.g., a light is either on or off), for which "additivity" is distinct from "invariance" as we explain later, the rest of our chapter concerns binary outcomes.

Terminology and background information

We first clarify some terminology and present some background information. In the previous section, we argued that causal relations are not present in the input to the cognitive system. Given that people do "know" causal relations, such knowledge of causality must therefore emerge somewhere along the pathway from the sensory input to its ultimate output. Under the modularity assumption in cognitive science (Marr, 1982), we designate the causal induction module, the focus of this chapter, to be the segment deep in the computational pathway that begins with an input layer, the layer closest to the causal output that is not yet causal. This module takes as input heterogenous noncausal information encompassing event frequencies and

variable intensities and generates as its output *causal representations* and judgments about them. The goal of the module is to induce useable causal representations.

We assume that, for the module, a *causal representation* consists of a cause with one or more component factors, an outcome, and the causal relation between the cause and the outcome. A cause (e.g., stormy weather) influences an outcome occurring in an entity (e.g., an airplane's safe landing). Causal influence can be generative (e.g., increasing the probability of a safe landing) or preventive (reducing that probability). Causal relations are asymmetric in that a cause brings about its outcomes, but an outcome does not necessarily bring about its causes. The temporal order of causal relata is asymmetric in that causes precede, or, in some cases, occur simultaneously with the outcomes they bring about (e.g., a wall's blocking of the sun causing the occurrence of a shadow on the ground), but outcomes never precede their causes.

The process of causal induction does its work in situations where available domain-specific causal knowledge does not favor whether or not the candidate is indeed a cause of the outcome (i.e., the input is noncausal in that sense), so that the resulting causal judgment regarding that relation is new knowledge. To be sure, causal knowledge can be transmitted from one reasoner to another (e.g., via verbal communication) after that knowledge has already been induced. However, causal knowledge is ultimately induced based on some individual's experience, through observations of particular events involving the states of candidate causes and of an outcome. (For counterexamples, see Garcia & Koelling, 1966, for evidence of causal knowledge that is not acquired due to individuals' experience, but via the process of evolution based on species' experience.)

We take the view that causation is represented as taking place in individual entities (token causation). But, because causation is "unobservable", no judgment can be made regarding what caused an outcome based on the state of the candidate cause and of the outcome in an individual entity. Causal induction therefore concerns causes and outcomes that are categories (type causation). They are categories in the sense that each is characterized by one or more properties that are common across multiple particular cause and outcome events. These categories are diverse, reflecting reasoners' concerns: They might represent some external event (e.g., rainy weather), some overt action (e.g., a person taking ibuprofen), some enduring state of an entity (e.g., oxygen being present on the surface of the Earth), or an entity's having or not having some

property (e.g., a person having or not having a headache), or some other kind of event. In other words, type causal judgments are inferences about relations between cause- and effect-categories based on observations of sets of events across time in which instances (tokens) of cause-categories variously present or absent in sets of entities are, with various probabilities, associated with instances (tokens) of an outcome category occurring in those entities (see Woodward, 2003, for a discussion of the relation between type and token causation; see Stephan & Waldman, 2018, this volume; and Stephan, Mayrhoer, & Waldman, 2020, for discussion of how reasoners can apply their generic, type-level, causal knowledge about causal strength to assess token or singular causation).

Prerequisites for evaluating the influence of a candidate cause on an outcome

No judgment about the influence of a candidate cause on a target outcome can be made based solely on what we term *cause-present* information – that is, information on the relative frequency with which the outcome occurs in multiple entities in which the candidate cause is present. This is so because causation is unobservable. The outcome could have occurred due to *background causes*, various other (known and unknown) causes of the outcome present in the context. The influence of the candidate cause therefore needs to be teased apart from that due to the background causes. Doing so requires an estimate of the probability of the outcome due to background causes.

The influence due to background causes in the cause-present events can be estimated, counterfactually, by the relative frequency of the outcome in *control events*, those in the same causal context— that is, with the same background causes present— but lacking the candidate cause (i.e., *cause-absent* events). In other words, to infer the relation between a candidate cause and an outcome, a reasoner relies on observation of two relative frequencies: the relative frequency with which an outcome occurs in cause-present events *and* that in control events. Situations in which background causes are held constant – where there is "no confounding" – are the only ones that license causal induction based on contrasting the probabilities of the outcome in the cause-present and cause-absent events (Cheng, 1997). The probability of the outcome in control events, due to satisfaction of the "no confounding" prerequisite, provides an estimate of the probability of the outcome in cause-present events (assuming that sample sizes are sufficiently large). With that estimate, it becomes possible for a reasoner to *decompose* the

occurrence of the outcome in the cause-present events into an estimate of the proportion brought about by the candidate cause and the proportion brought about by the background causes, with a potential overlap between the two subsets of events. Given that causation is never observable, decomposition is essential to causal induction.

In the following, we will go through an example where a medication taken by human patients constitutes the candidate cause of headache, the target binary outcome. We partition all causes of headache into the candidate cause and a *composite* of all other (known and unknown) causes in the context (i.e., the background causes), which may affect the occurrence of headache independently of the medication or interacting with the medication. To use the headache example, the composite of background causes might include stress, dehydration, or sleep deprivation.

Because causes and effects are categories, causal induction involves hypothesizing and evaluating representations of a candidate cause, a process that may be parallel to the process involving judgments about causal structure and causal strength (Kemp, Goodman, & Tenenbaum, 2010; Lien & Cheng, 2000; Marsh & Ahn, 2009; Waldmann & Hagmayer, 2006; Waldmann, Meder, Sydow, & Hagmeyer, 2010). This is an important aspect of causal induction which we do not address in the present chapter.

Estimates of causal strength depend on the assumed decomposition function

This process of decomposing the probability of a binary outcome into contributions by its various causes to estimate the causal strength of the candidate can be formally specified using a *decomposition function*. Importantly, given observations of the same event frequencies, different decomposition functions yield different estimates of a candidate's causal strength. This divergence between commonly considered decomposition functions in the psychological literature does not pertain to continuous outcomes (e.g., a light can have varying degrees of brightness), because the dominant decomposition function, additivity, is the causal-invariance function for continuous outcomes.[3] In the following, we focus on causal events featuring a binary

[3] Different outcome-variable types (e.g., binary, continuous, vectors, waves) have different causal-invariance functions, depending on how superposition of causal influences (i.e., independent influences) is normatively expressed in mathematical form. Vector addition, for example, is the causal-invariance function for vectors.

outcome in order to contrast two decomposition functions: 1) the *causal invariance* decomposition function, and 2) an associative foil that we label the *additive* decomposition function. Empirical evidence comparing the two decomposition functions shows that the former but not the latter function is descriptive of human causal induction (e.g., Buehner, Cheng, & Clifford, 2003; Cheng, 1997; Liljeholm & Cheng, 2007; Lu, Yuille, Liljeholm, Cheng, & Holyoak, 2008).

Our goal in contrasting these two decomposition functions is to demonstrate that analytic knowledge of causal invariance, in the form of the causal-invariance decomposition function applied to a candidate cause and the composite of other causes in the context, is a rational constraint on causal induction. We show that use of only the causal-invariance decomposition function during learning will result in a logically consistent indication of whether the target causal relation indeed generalizes to other contexts, judging by the criterion of *symmetry*: a causal relation that generalizes from a learning context to an application context should also generalize in the reverse direction, from the application context back to the original learning context.

Please note that the extended example to follow makes use of headache as a *binary outcome*. We acknowledge that headaches do, indeed, vary in their intensity and are more realistically understood as a continuous outcome. Our reason for using headache in our example is simply that their presence and absence is easy to represent in visual diagrams. For more realistic examples of a binary outcome, consider outcomes such as a woman being pregnant or not, a reader subscribing to a magazine or not, a car's motor being on or off, an organism being alive or dead, or a protestor infected with COVID-19 or not.

Let us consider the following situation, which we will call Context 1: the candidate cause is the medication M taken by human patients and the outcome is these patients having a headache. Patients are randomly assigned to two groups: one that received medication M, another that does not. No relevant causal knowledge about individual patients is available. Here, it is important to note that *individual patients* are the meaningful units within which the medication exerts its causal influence.

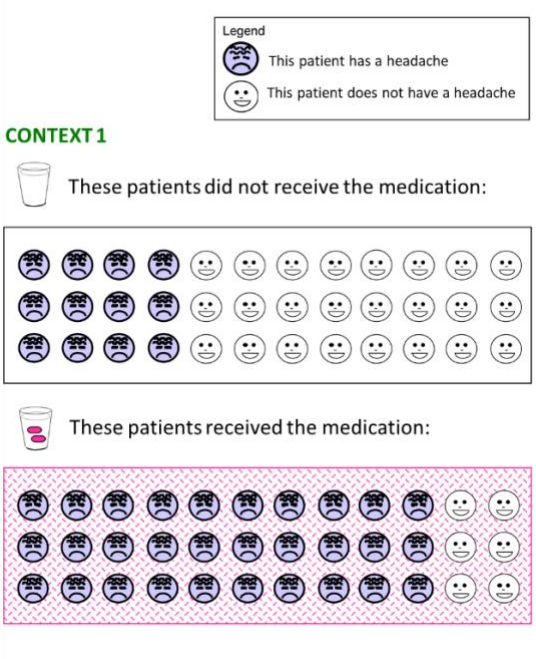

Figure 5. Occurrences of headaches in patients who did not receive the medication (top) and in patients who did receive the medication (bottom) in Context 1.

Figure 5 depicts the occurrence of headaches in patients who did not (top panel) and who did receive the medication (bottom panel) in Context 1. As the figure shows, when patients each take the medication (i.e., in cause-present events), 30/36 of them develop a headache, and when patients do not take the medication (i.e., in control events), 12/36 of them develop a headache. (We leave the fractions unreduced so that they readily correspond to the relative frequencies in the figures.) These relative frequencies of headache are best understood on a ratio scale – each expresses the proportion of individual patients who has a headache relative to the entire group of patients in each kind of event. Assume that there are no preventive causes. When considered in an experimental setting, each of these cause-present and control events might be considered an experimental trial.

Let us consider the causal strengths inferred by the two decomposition functions. An *additive* decomposition function represents the probability of patients having a headache (H) after having taken the medication M in the cause-present events ($P(H = 1|M = 1, B = 1) = 30/36$) as the sum of (1) an estimate of the probability across patients that headache occurs attributable

to[4] the composite of background causes B (p_B = 12/36) and (2) an estimate of the probability that taking the medication brings about a headache across patients ($p_{M_{Additive}}$ = 18/36):

$$P(H = 1 \mid M = 1, B = 1) = p_B + \boldsymbol{p_{M_{Additive}}} \quad (1)$$
$$\frac{30}{36} = \frac{12}{36} + \frac{\mathbf{18}}{\mathbf{36}}$$

The additive decomposition function instantiates an exertion of causal strength where each patient is only susceptible to developing a headache from *either* background causes *or* the medication, but not both. To put the point differently, the additive decomposition function implies that in events where the background causes exert their causal strength, the medication *withholds* exerting its causal strength, and in events where the medication exerts its causal strength, the background causes *withhold* exerting their casual strength. Such an absurd state of affairs would involve the medication and the background causes *knowing* in which patients each other causes headaches and having the ability to *control* when they themselves do so. In other words, the medication and the background causes are not acting independently.

On the other hand, a *causal invariance* decomposition function represents the probability of this same cause-present outcome (P(H = 1|M = 1, C = 1) = 30/36) as specified in Eq. 2: as a superposition of the independent influences of the medication and the background. $\boldsymbol{p_{M_{Invariance}}}$ in Eq. 2 is the causal power of the medication. *Causal power* is a theoretical, unobservable probability which represents the capacity for an instance of a cause in an entity to bring about an instance of an outcome in that entity (Cartwright, 1989; Cheng, 1997). In the absence of relevant causal knowledge about the individual entities exposed to the candidate cause, the induction of causal power of the candidate based on observations of the state of the cause and of the outcome in a set of entities is constrained by the default assumption that the power of the candidate is *independently and identically distributed* (*iid*) across those entities (e.g., Casella & Berger, 2001). Each particular instantiation of a given cause in an entity is assumed to independently exert the *same* causal power to bring about an instantiation of its outcome across all entities in

[4] We distinguish the interpretation of p_B mentioned in the text from an alternative interpretation in which it is the probability with which background causes bring about headache. The latter is not estimable because of the inherent lack of information about the probability of the occurrence of unobserved and unknown background causes in the context.

the set exposed to the same candidate cause. There is no reason to assume otherwise in the absence of relevant causal knowledge. The independent exertion of causal power across individual patients is captured in the intuition that the medication in one patient does not know what the medication in another patient does.

The terms on the right-hand side of Eq. 2 are respectively: (1) an estimate of the probability that headache occurs across patients attributable to background causes (p_B = 12/36), (2) an estimate of the probability that taking the medication brings about a headache across patients ($p_{M_{Invariance}}$= 27/36), and (3) the counterfactual probability that headache would be produced by taking the medication if it had not already occurred due to the background causes ($p_{M_{Invariance},B}$ = 9/36), estimated by the product of the preceding two terms:

$$P(H = 1 \mid M = 1, B = 1) \;=\; p_B + \boldsymbol{p_{M_{Invariance}}} - p_{M_{Invariance},B} \qquad (2)$$

$$\frac{30}{36} = \frac{12}{36} + \frac{\mathbf{27}}{\mathbf{36}} - \frac{9}{36}$$

The causal invariance decomposition function arrives at $p_{M_{Invariance}}$ = 27/36 as the causal power of taking medication. Under this interpretation, every patient taking the medication in this context is just as susceptible as any other patient to develop headache from background causes, and, independently, likewise from taking the medication. This means that there are 9/36 cases in this context where patients' experiencing relief from headache is *causally overdetermined*. Those are the cases in which the background causes and taking medication are *independently* sufficient to cause headache such that, counterfactually, the absence of either one would still have resulted in headache. Those cases are represented by the $p_{M_{Invariance},B}$ term (which is subtracted in accordance with probability theory to avoid counting those cases twice).

It should be clear that the additive decomposition's estimates of causal strength (i.e., of background causes and of taking medication on developing headaches) *violate* the iid condition across patients. Whereas we refer to the estimate that instantiates the iid assumption as causal power, we use *causal strength* as the theoretically neutral term when an estimate does not necessarily instantiate the iid assumption. We therefore refer to $p_{M_{Additive}}$ as a *causal strength* estimate.

Our present aim is to explain why this assumption is a rational constraint on inducing useable causal representation. To gain an intuitive sense of the superposition, consider: What causal strength of medicine M would most likely result in the outcome depicted in the experimental group in Figure 5 (the bottom panel) —assuming that M and the background do not interact —if patients in the control group in that figure (the top panel) had received medicine M? We hope it is intuitive that the answer is 3/4, the maximum-likelihood estimate of medicine M producing headache (Griffiths & Tenenbaum, 2005) under the assumptions of the causal power theory (Cheng, 1997).

Are causal strengths inferred without the iid assumption generalizable?

Thus, we see how different decomposition functions arrive at different estimates of causal strengths. To illustrate that causal strengths that violate the iid assumption would not be usable causal knowledge, we explore generalization to a different causal context, which we will call Context 2. Context 2 is the *application context* to which we apply the causal strengths inferred in Context 1, the *learning context*. We continue with the same candidate cause (i.e., taking medication) and the same target outcome (i.e., experiencing headache) as Context 1. We show that, unlike $p_{M_{Invariance}}$, $p_{M_{Additive}}$ does not satisfy even the minimal generalization requirement: specifically, after $p_{M_{Additive}}$ successfully generalizes to Context 2, it fails to generalize from Context 2 back to Context 1, the original context in which it was inferred. To illustrate that $p_{M_{Additive}}$ fails to satisfy this minimal requirement, we chose Context 2 to be a situation in which all associative and causal models agree on the predicted outcome from introducing a cause with any given strength. In Context 2, 0/36 patients have a headache without any medication (see Figure 6), indicating that there are no background causes, so that any candidate cause introduced is the only cause present. As before, assume that there are no preventive causes.

An *integration* function uses estimates of causal influence induced from prior experience (e.g., from observing event frequencies in Context 1) to predict frequencies of some outcome to a new context (e.g., the occurrence of headache in patients having taken the medication in Context 2). Importantly, an integration function *generalizes* an estimate of causal strength to a novel context, in which we have no information yet on whether and how the background causes in that context interact with the target cause. Without any such information, the only reasonable default

assumption is that the target cause brings about the outcome of interest with the *same* capacity on each event, and this assumption is captured by the iid nature of causal power. In other words, an integration function that instantiates this iid nature of causal power is the only justifiable integration function to apply as a default. An integration function assuming iid specifies the inverse operation as the causal-invariance decomposition function. We use this function below to generate predictions of headache occurrence in Context 2, for both $p_{M_{Invariance}}$ *and* $p_{M_{Additive}}$. Because medication M is the only cause in Context 2, no superposition is involved. The applications of this integration function we illustrate below are therefore trivial, and the resulting predictions do not differ from those resulting from applying an additive integration function.

Recall that the causal strength estimate from the additive decomposition function in Context 1 was 18/36 and that the causal power estimate from the causal invariance decomposition function was 27/36. Also recall that 0/36 patients develop a headache without the medication in Context 2. Incorporating this outcome frequency with the causal strength estimated by the additive decomposition function in Context 1 yields the prediction that 18/36 patients will develop a headache after taking the medication:

$$p_B + p_{M_{Additive}} - p_{M_{Additive},B} = \boldsymbol{P(H = 1 \mid M = 1, B = 1)_{Additive}} \quad (3)$$

$$\frac{0}{36} + \frac{18}{36} - \frac{0}{36} = \mathbf{\frac{18}{36}}$$

And doing the same but instead using the causal power estimated by the causal-invariance decomposition function in Context 2 yields the prediction that 27/36 patients will develop a headache after taking the medication:

$$p_B + p_{M_{Invariance}} - p_{M_{Invariance},B} = \boldsymbol{P(H = 1 \mid M = 1, B = 1)_{Invariance}} \quad (4)$$

$$\frac{0}{36} + \frac{27}{36} - \frac{0}{36} = \mathbf{\frac{27}{36}}$$

We now compare how the predictions generated using the two causal strength estimates generalize the respective strengths back to Context 1, the learning context.

Across the causal events to which the same causal representation applies, an agent may use their observations of some causal events to *induce* that causal representation, or they may

apply their causal representation to either explain or predict outcome occurrences in other causal events. We refer to the cognitive processes underlying the former phenomenon as *causal induction* and those underlying the latter phenomenon as *causal reasoning*. While both processes may operate in a given event, for the purpose of exposition it is worthwhile to distinguish between *learning contexts* where an agent engages in causal induction and *application contexts* where an agent engages in causal reasoning.

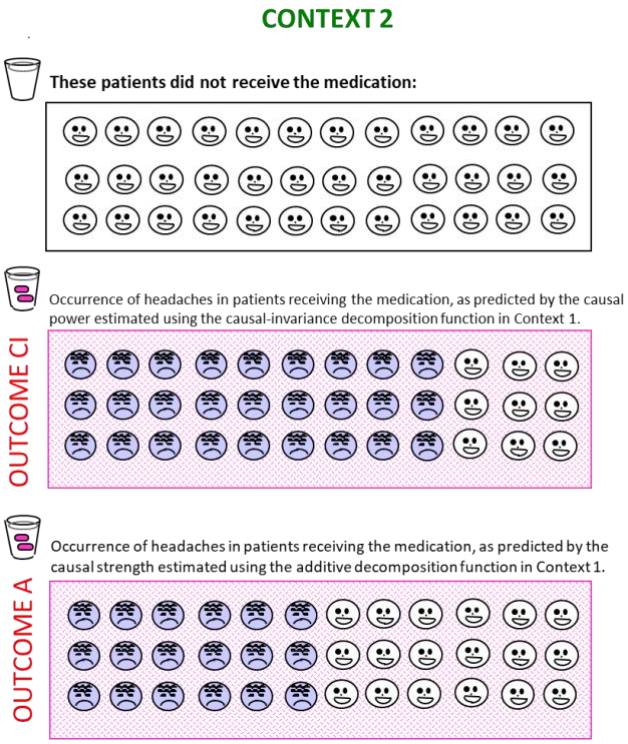

Figure 6. Occurrence of headaches in patients who did not receive the medication in Context 2 (top) and occurrence of headaches in patients receiving the medication in this context as predicted, respectively, by the causal power estimated in Context 1 using the causal-invariance decomposition function (middle) and by the causal strength estimated in Context 1 using the additive decomposition function (bottom).

Even though the distinction between the learning and application contexts is natural with respect to a reasoner's cognitive history, this same distinction is completely incidental with respect to *rationally generated* causal representations. Specifically, if the same causal representation holds across two causal contexts, which of those contexts was a learning context

and which was an application context for a particular reasoner should make no difference. Returning to our example above, while Context 1 served as a learning context and Context 2 served as an application for a hypothetical reasoner, reversing their roles (i.e., so that Context 2 serves as a learning context and Context 1 serves as an application context) should yield consistent inferences. In other words, causal induction should accommodate symmetry between learning and application contexts. It is logically inconsistent for a causal strength to be *both* "the same" and "not the same" across two contexts: if a cause operates the same way in an application context as in its learning context, its causal strength should remain the same across the two contexts, regardless of that context's epistemic relation to the reasoner. In the following, we show that causal induction assuming causal invariance as the decomposition function does accommodate this symmetry, but that assuming the additivity decomposition function or any other non-causal invariance decomposition function fails to accommodate this symmetry. Let us now flip the learning and application contexts of our previous example, treating Context 2 as a learning context and Context 1 as an application context.

Figure 7 depicts the results of flipping the learning and application contexts for the additive and causal invariance functions. The causal strength estimated by the additive decomposition in Context 2, $p_{M_{Additive}}$= 18/36, makes an incorrect prediction that 24/36 of patients in Context 1 will have headache after having taken the medication:

$$p_B + p_{M_{Additive}} - p_{M_{Additive},B} = \boldsymbol{P(H = 1 \mid M = 1, B = 1)_{Additive}}$$
$$\frac{12}{36} + \frac{18}{36} - \frac{6}{36} = \mathbf{\frac{24}{36}}$$

And as should be obvious, because of the inverse relation between decomposition and integration, the causal power estimated by the causal invariance decomposition function in Context 2, $p_{M_{Invariance}}$= 27/36, makes the correct prediction that 30/36 of patients in Context 1 will have headache after having taken the medication:

$$p_B + p_{M_{Invariance}} - p_{M_{Invariance},B} = \boldsymbol{P(H = 1 \mid M = 1, B = 1)_{Invariance}}$$
$$\frac{12}{36} + \frac{27}{36} - \frac{9}{36} = \mathbf{\frac{30}{36}}$$

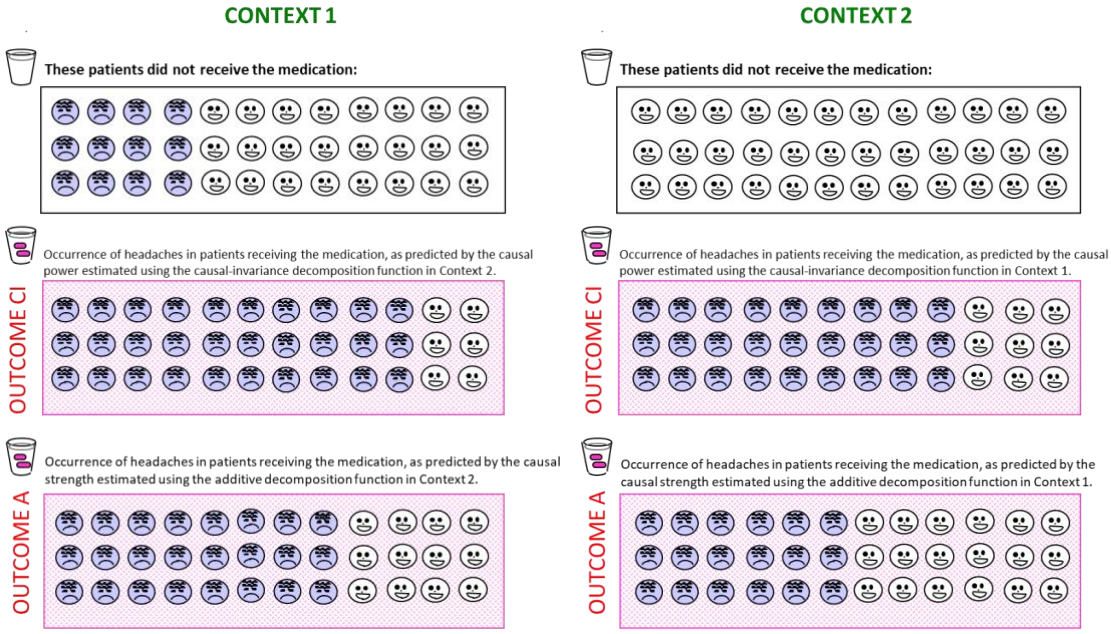

Figure 7. Occurrence of headaches in patients who did not receive the medication in Context 1 (top left) and occurrence of headaches in patients receiving the medication in this context as predicted, respectively, by the causal power estimated using the causal-invariance decomposition function (middle left) and by the causal strength estimated using the additive decomposition function (bottom left) based on the event frequencies in Context 2 shown in Figure 6 and duplicated here for easier visual comparison (top right).

Here, we see that generalizing from Context 2 to Context 1 using the causal invariance decomposition function, but not the additive decomposition function, accommodates the symmetry between learning and application contexts. Specifically, the predicted occurrence of headaches in patients taking the medication in Context 1 using the causal power estimated by the causal-invariance decomposition function in Context 2, but not that estimated by the additive decomposition function, yields the actual observed frequency in Context 1 (Figure 5). This difference in the satisfaction of the symmetry requirement follows from 1) the inverse relation between a decomposition function and an integration function, and 2) the inherent assumption when a reasoner applies causal knowledge to a new context where potentially different unobserved or unknown causes occur: the causal relations being generalized operate the same

way across the learning and application contexts. That inherent assumption renders the causal-invariance function the only rational integration function to apply in causal reasoning. As an example, we have shown that the additive decomposition function fails to be logically consistent across transpositions of the arbitrary "learning" and "application" context labels for inducing causal relations involving a binary outcome, illustrating that the causal-invariance function is the only rational decomposition function to apply in causal induction.

A careful reader may notice that the additive decomposition function that we have discussed thus far is the one that underlies a linear regression model. Considering that our example features a binary outcome, such a reader might protest that it is inappropriate to use linear regression to estimate binary outcomes (but see Gomilla, 2020 for advocation for this very practice) and question the relevance of problematizing the additive function for such situations. In the following, we show that the inconsistency described above is also characteristic of the logistic model, whose use in predicting binary outcomes is much more conventional. By extending our analysis to the logistic model, we argue that this logical inconsistency is inherent to any decomposition function that violates the iid assumption (i.e., any non-causal-invariance function).

Is logical inconsistency a problem for generalized linear models?

The logical inconsistency across contexts is true not only of the additive decomposition function, for which this problem may be obvious. Here we illustrate the general problem with a concrete example where it is easy to see the problem for the logistic function, a generalized linear function:

$$f(z) = \frac{e^z}{1+e^z}, \tag{5}$$

where *z* may be interpreted as the weighted sum of the predictor variables (in this interpretation the causal variables), and $f(z)$ is the probability of the binary outcome in question. The logistic function (Eq. 5) is assumed by "normative" associative models such as logistic regression, a widely used statistical method in medical and business research, where binary outcomes (e.g., a tumor is either malignant or benign, a bone is either fractured or intact) are common.

One interpretation of the logistic model is that: 1) the predictor variables exert their independent influences, not directly on the binary outcome, but on a latent mediating variable *s* —a continuous variable with values on an interval or ratio scale— so that *s* is a weighted sum of the predictor variables, 2) the probability *f*(*z*) in Eq. 5 can be conceptualized as being due to noise *n* being added to the latent variable *s* to produce a decision *y* representing the binary outcome; *n* has a logistic distribution (i.e., density function) with a mean of 0 and a scale parameter equal to 1, and 3) when $s + n$ is greater than a threshold—0 in this case, the binary outcome occurs, otherwise the outcome does not occur; that is:

$$y = \begin{cases} 1 & s + n > 0 \\ 0 & \text{else.} \end{cases}$$

We do not dispute that hypotheses with a mediating variable should be considered. However, the principle of parsimony would be violated if the continuous mediating variable is postulated as a default, bypassing consideration of a simpler hypothesis. In cases in which the simpler independent-influence hypothesis is in fact the better explanation, it would never be found. For this reason, the common usage of logistic regression as the standard statistical method for analyzing data with a binary outcome is likely to have contributed to the replicability crisis (Ioannidis, 2005; Open Science Collaboration, 2015).

A shared mediating variable can strain credulity in some cases. Consider one of the binary outcomes introduced briefly earlier: Pregnancy. Pregnancy is likely to have dissociable, independent causes. Two such causes might include, 1) whether or not someone has received a medical procedure to improve their fertility and 2) whether or not someone lives in a country with a policy that limits child-rearing (e.g., China and its one-child policy). The mechanism by which someone having received a medical procedure to improve their fertility influences their chances of getting pregnant is biological and internal to their bodily function. On the other hand, the mechanism by which someone's living in a country with a policy that limits child-rearing is social and external to their bodily function. A latent variable that elides the clear distinction between these two causal mechanisms seems implausible. To what would this latent variable refer?

It is our understanding that the simpler hypothesis, the independent *direct* causal-influences hypothesis without the continuous mediating variable, is typically not in the repertoire of potential models to evaluate in popular statistical-analysis software (e.g., SPSS, R). The software user has no choice but to posit the more complex hypothesis, which implies foregoing deviation from independent direct causal influences as a criterion for hypothesis revision, instead treating independent influences on the continuous intervening variable as the aspiration (Bye, Chuang, & Cheng, under review). Beyond the logistic model's relative lack of parsimony, it fails to estimate causal strengths that generalize across distinct contexts when the simpler hypothesis holds, as we now move on to show.

To see the problem with the logistic function as a decomposition function for sets of events with a binary outcome, let us consider yet another context, which we will call Context 3, alongside the previously discussed Context 1 (see Figure 8). In Context 3, 6/36 patients develop a headache without receiving the medication, and 24/36 patients develop a headache after having taken the medication.

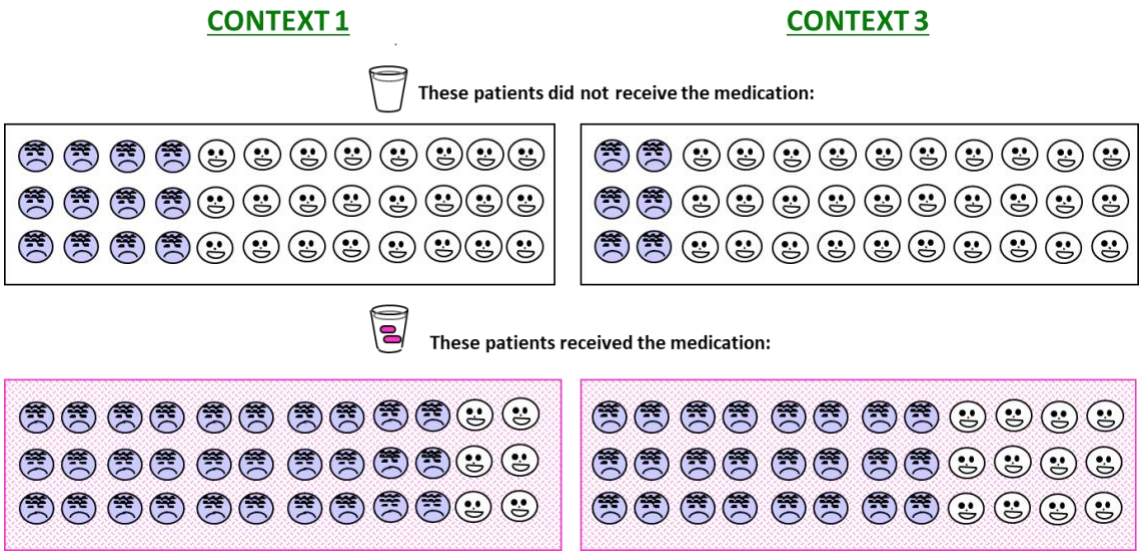

Figure 8. Occurrences of headaches in patients who did not receive the medication in Context 1 (top left) and Context 3. Occurrences of headaches in patients receiving the medication in Context 1 (bottom left) and in Context 3 (bottom right).

We constructed the outcome frequency for control events in Context 1 to be the complement of the outcome frequency for cause-present events in Context 3, and the outcome frequency for control events in Context 3 to be the complement of the outcome frequency for control events in Context 1. We use this complementary pattern as an obvious example to illustrate the general violation of the iid assumption by this model. In this model, the probability of a binary outcome is a logistic function of *z*, the weighted sum of the predictor variables, the causal variables in the case of our example:

$$z = B_1 * w_{B_1} + M * w_M \quad (6)$$
$$B_1 \in \{0, 1\}$$
$$M \in \{0,1\}$$

B_1 refers to *Context 1*, and *M* refers to the medication. Each are binary variables where a value of 1 represents the presence of their referent, and a value of 0 represents its absence. w_{B_1} and w_M represent the causal weights associated with the *Context 1* background causes and the medication, respectively. w_{B_1} is the same as p_B from the additive decomposition function applied to the respective causal context. Decomposing using Eqs. 5 and 6, the logistic function, like the additive decomposition function, characterizes the medication as an invariant cause across Context 1 and 3, in that w_M is constant across these contexts. It should be clear that this characterization is mistaken. For each of these contexts, consider: What causal strength of medicine M would most likely result in the outcome depicted in the experimental group in Figure 8 (the bottom panel) —assuming that M does not interact with the background in either context —if patients in the control group in that figure (the top panel) had received medicine M? The answers are not the same for the two contexts. They instead correspond to $p_{M_{Invariance}}$ in Context 1 and in Context 3.

We will now explain the general divergence between the logistic decomposition function and the causal invariance decomposition function by examining the logistic function graphically, as shown in Figure 9. Each of the four pairs of light grey and dark grey points represents a different causal context. Here, we see that the four depicted causal contexts that are symmetric about *z* = 0 or *f(z)* = .5 (e.g., the two contexts represented by the two inner pairs of points or the

two contexts represented by the two outer pairs of points in the figure), w_M will be identical across contexts, and the medication will be represented as an invariant cause. This explains the point made earlier: *Context 1* and *Context 3* in Figure 8 were constructed to be symmetric about *z* = 0 or *f(z)* = .5. (But note that Figure 9 does not illustrate Contexts 1 and 3.)

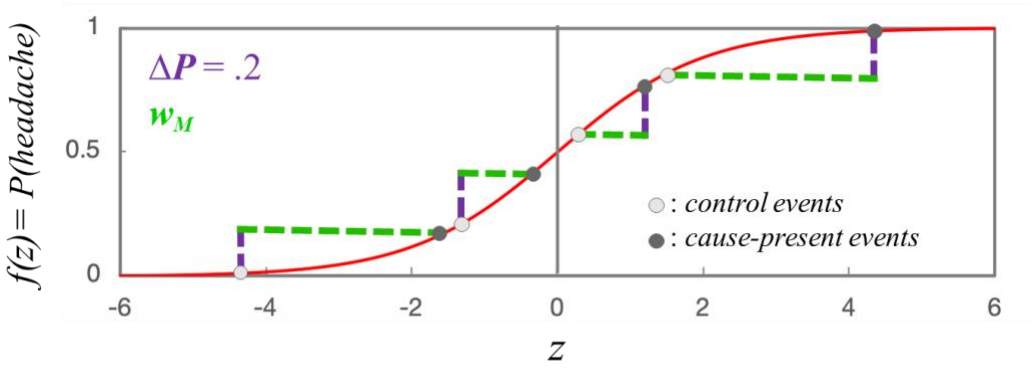

Figure 9. w_M according to the logistic decomposition function in various causal contexts. Each pair of cause-present and control events (light grey and dark grey points) represents a different causal context.

Let us now shift attention to the whole logistic curve. Notice that each causal context in Figure 9 has the same *ΔP* (vertical dashed lines, purple in online version), that is, the *difference* in headache probability between cause-present events and the control events is held constant across contexts. In other words, *P(H = 1|B = 1,M = 1)* – the probability of headache observed in cause-present events – is equal to the sum of *ΔP* and *P(H = 1|B = 1, M = 0)* – the probability of headache observed in control events. Let us now focus on w_M (horizontal dashed lines, green in online version) across contexts. Notice that, for causal events with the same *ΔP*, as the observed headache probability for control events (light grey points) approaches .5, w_M decreases, and that as the observed headache probability for control events increases beyond .5, w_M increases. This trend is shown in the left panel of Figure 10, which depicts causal strength estimate from the logistic decomposition function, w_M (horizontal dashed lines in Figure 9, green in online version), as a function of headache probability in control events (*y*-values of the grey dots in Figure 9), holding *ΔP* constant. For comparison, the right panel of Figure 10 depicts the causal power estimate from the causal invariance decomposition function, $p_{M_{Invariance}}$, as a function of headache probability in control events, holding *ΔP* constant. In sharp contrast to w_M, $p_{M_{Invariance}}$ monotonically increases as the outcome probability in control

events increases. Intuitively, this is because increases in the outcome probability in control events, by counterfactual reasoning, imply a larger *proportion* of patients who would have been *without* headaches in the cause-present events who are caused to have headache, as expressed by $p_{M_{Invariance}}$.

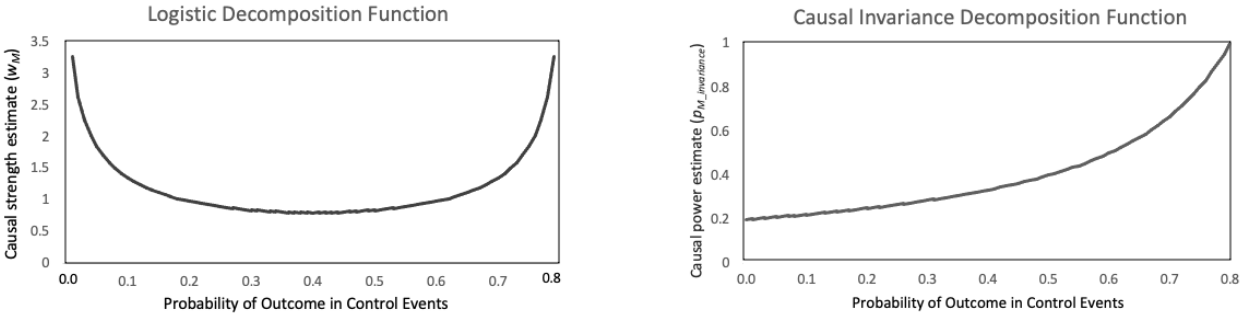

Figure 10. Causal estimates (*y*-axes) as a function of outcome probability in control events (*x*-axes), holding constant ΔP = .2.

In showing this divergence between the logistic decomposition function and the causal invariance decomposition function across Contexts 1 and 3, we have demonstrated that generalized linear models such as logistic regression are logically inconsistent with causal generalization, and do not yield usable causal knowledge. Further, in showing their incompatibility with causal invariance, we have shown that estimates of causal strength that are consistent with generalized linear models diverge from human causal induction (Buehner, et al., 2003; Cheng, 1997; Liljeholm & Cheng, 2007; Lu et al., 2008).

Through discussing event frequencies across Context 1, 2, and 3, we have argued for causal invariance as a rational constraint on the formulation of useable causal knowledge. Causal invariance assumes the identical and independent exertion of causal power within and across causal contexts. Specifically, we have shown 1) how different quantitative estimates of causal strengths assuming non-causal-invariance and causal-invariance decomposition functions respectively violate and accommodate this constraint, and 2) violation of the constraint leads to logical inconsistency, resulting in false alarms and misses in the hypothesis testing and hypothesis revision process. The violation is not specific to the cases we illustrated, but inherent to non-causal-invariance functions.

In conclusion, revisiting Hume's (1739/1987) radical insight that causation is "unobservable", we see that it is strongly supported by findings and theoretical developments from diverse perspectives. These perspectives — psychological science, physics, vision science, and neuroscience — converge in clarifying that causation is a representation of the empirical world by and in our mind. Grounded in a representation-dependent conception of reality, we see that the unobservability of causal relations demands an explanation of how useable causal knowledge is attainable in a search for such knowledge within an infinite space of possible representations. Our analysis shows that it is attainable only if a cognitive process adopts causal invariance as the default decomposition function, in other words, implements the assumption that there exist invariant causal relations in the world, and (implicitly) aims to construct such knowledge. This assumption narrows the search space to representations that are 1) candidates for serving our species' subjective goal of possessing useable causal knowledge and 2) logically consistent with that goal. Omitting the constraint results in logical inconsistency during the search in that vast space. It follows that "normative" statistical inference, whether frequentist or Bayesian, will not yield useable causal knowledge if it violates the causal invariance constraint.

Acknowledgments

We thank Hongjing Lu and Simon Stephan for very helpful comments on a draft of this chapter. We thank George Sperling for very helpful discussion.

References

Armel, K. C., & Ramachandran, V. S. (2003). Projecting sensations to external objects: Evidence from skin conductance response. *Proceedings of the Royal Society of London B, 270*, 1499-1506.

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

Brasil-Neto, J. P., Pascual-Leone, A., Valls-Sole, J., Cohen, L. G., & Hallett, M. (1992). Focal transcranial magnetic stimulation and response bias in a forced-choice task. *Journal of Neurology, Neurosurgery, and Psychiatry*, *55*, 964-966.

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. (under review). How do causes combine their effects? The role of causal invariance for generalizable causal knowledge. *Cognition*.

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

Casella, G., & Berger, R. L. (2002). *Statistical Inference*. Duxbury Press.

Cheng, P.W. (1993). Separating causal laws from causal facts: Pressing the limits of statistical relevance. In D. L. Medin (Ed.), *The Psychology of Learning and Motivation*, Vol. 30 (pp. 215-264). New York: Academic Press.

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

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 Waldmann, M. R. (Ed.), *The Oxford Handbook of Causal Reasoning*. New York: Oxford University Press.

Garcia, J., & Koelling, R. A. (1966). Relation of cue to consequence in avoidance learning. *Psychonomic Science, 4*(3), 123–124. https://doi.org/10.3758/BF03342209

Gomilla, R. (2020). Logistic or linear? Estimating causal effects of treatments on binary outcomes using regression analysis. *Journal of Experimental Psychology: General*.

Goodman, N. (1955). *Fact, Fiction, and Forecast*. Cambridge, Massachusetts: Harvard University Press.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*, 354-384.

Hawking, S., & Mlodinow, L. (2010). *The grand design*. New York: Bantam Books.

Henle, M. (1962). On the relation between logic and thinking. *Psychological Review, 69*(4), 366-378.

Hofer, H., Singer, B. & Williams, D.R. (2005). Different sensations from cones with the same photopigment. *Journal of Vision*, ***5***, pp. 444-454.

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

Hume, D. (1748/1975). *An Enquiry Concerning Human Understanding*. (L. A. Shelby-Bigge & P. H. Nidditch, Eds.) (3rd ed.). Oxford: Clarendon Press.

Ioannidis, J.P.A. (2005). Why most published scientific findings are false. *PLoS Medicine*, *2*(8), e124.

Kant, I. (1781/1965). *Critique of Pure Reason*. London: Macmillian.

Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2010). Learning to learn causal models. *Cognitive Science*, *34*(7), 1185-1243..

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

Liljeholm, M., & Cheng, P. W. (2007). When is a cause the "same"?: Coherent generalization across contexts. *Psychological Science*, *115*(4), 955-984.

Lu, H., Yuille, A., Liljeholm, M., Cheng, P.W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. *Psychological Review*, *18*(11), 1014-1021.

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

Marsh, J. K. & Ahn, W. (2009). Spontaneous assimilation of continuous values and temporal information in causal induction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 334-352.

Michotte, A. (1946/1963). *The Perception of Causality*. New York: Basic Books.

Mitchell , D.E. & Rushton, W.A.H. (1971). Visual pigment in dichromats. *Vision Research, **11***, pp. 1033-1043.

Newton, I. (1687/1713/1726/1999). *The Principia: Mathematical Principles of Natural Philosophy*. (I. B. Cohen & A. Whitman, Eds.). Berkeley and Los Angeles: University of California Press.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231-259.

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

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.

Rips, L. J. (2011). Causation from perception. Perspectives on Psychological Science, 6: 77-97.

Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences*, *4*(8), 299-309.

Stephan, S., & Waldmann, M. R. (2018). Preemption in singular causation judgments: A computational model. *Topics in Cognitive Science*, *10*(1), 242-257.

Stephan, S., Mayrhofer, R., & Waldmann, M. R. (2020). Time and singular causation – a computational model. *Cognitive Science*, *44*(7), e12871.

Waldmann, M. R., & Hagmayer, Y. (2006). Categories and causality: The neglected direction. *Cognitive Psychology*, *53*, 27-58.

Waldmann, M. R., Meder, B., von Sydow, M., & Hagmeyer, Y. (2010). The tight coupling between category and causal learning. *Cognitive Processing*, *11*, 143-158.

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

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation (Oxford Studies in the Philosophy of Science)*. Oxford, England: Oxford University Press.