Running Head: EVENT HIERARCHIES AND CAUSES

Causal inference and the hierarchical structure of experience

Samuel G. B. Johnson & Frank C. Keil

Department of Psychology, Yale University

Corresponding author:

Samuel Johnson
2 Hillhouse Ave.
New Haven, CT 06520
samuel.johnson@yale.edu

Accepted version, 28 October 2014

In press at *Journal of Experimental Psychology: General*
doi: 10.1037/a0038192

Abstract

Children and adults make rich causal inferences about the physical and social world, even in novel situations where they cannot rely on prior knowledge of causal mechanisms. We propose that this capacity is supported in part by constraints provided by *event structure*—the cognitive organization of experience into discrete events that are hierarchically organized. These event-structured causal inferences are guided by a *level-matching principle*, with events conceptualized at one level of an event hierarchy causally matched to other events at that same level, and a *boundary-blocking* principle, with events causally matched to other events that are parts of the same superordinate event. These principles are used to constrain inferences about plausible causal candidates in unfamiliar situations, both in diagnosing causes (Experiment 1) and predicting effects (Experiment 2). The results could not be explained by construal level (Experiment 3) or similarity-matching (Experiment 4), and were robust across a variety of physical and social causal systems. Taken together, these experiments demonstrate a novel way in which non-causal information we extract from the environment can help to constrain inferences about causal structure.

Keywords: Causal reasoning, Event representation, Time, Explanation, Prediction

People make their way in the world despite an extraordinarily complex environment, and sparse information that underdetermines the environment's true structure. Event representations constitute one tool we use to guide us through this complexity, helping us to keep track of what happens and shaping our concepts of the past and present. Causal knowledge is a second sense-making tool, enabling us to evaluate the plausibility of specific causal relations and allowing us to make predictions of, and interventions on, future events. Here, we investigate the link between these two sense-making tools, arguing that event representations provide multiple constraints on causal inference and that they do so across a wide range of specific contents in both the physical and interpersonal spheres.

Causal explanations are foundational to many of our beliefs and inferences about the world (Lombrozo, 2010). They shape our social predictions (Dunning, Griffin, Milojkovic, & Ross, 1990), impressions of others (Kunda & Thagard, 1996), and attributions of blame and punishment (Alicke, 1992; Cushman, 2008; Lagnado & Channon, 2008). We use causal beliefs in reconstructing our past (Hastie, 1984; Wells, 1982), making decisions (Sloman & Hagmayer, 2006), and solving problems (Cheng & Holyoak, 1985). Given the variety and importance of causal knowledge, people use a variety of strategies for *evaluating* causal explanations (Johnson & Keil, 2014), including statistical co-occurrence (Cheng, 1997; Griffiths & Tenenbaum, 2005), mechanistic plausibility (Fugelsang, Stein, Green, & Dunbar, 2004; Johnson & Ahn, *in press*) and structural features of explanations such as simplicity (Johnson, Jin, & Keil, 2014; Lombrozo, 2007) and scope (Johnson, Johnston, Toig, & Keil, 2014; Johnson, Rajeev-Kumar, & Keil, 2014; Khemlani, Sussman, & Oppenheimer, 2011; Read & Marcus-Newhall, 1993).

Before we can evaluate a causal explanation, however, we must first *generate* it. The computational challenges here are forbidding. Suppose, for example, that you notice your

colleague Diane is often late to work. What possible explanations might you generate for her lateness? Some possible causes—an unreliable alarm, chronic car problems—seem worthy of consideration, while other possible causes—an elaborate conspiracy, a moth's flapping wings in India—seem unworthy. It is not that we rapidly consider such implausible hypotheses and dismiss them—rather, *we never consider them at all*. We must therefore have a set of heuristics for generating potential causal candidates—candidates which may subsequently be rejected as implausible, but which are nonetheless worthy of consideration. Among the first to identify this hypothesis generation problem was the philosopher Charles Sanders Peirce (1997/1903), who noted that any observed evidence is consistent with an infinite number of potential hypotheses—an observation we refer to as *Peirce's problem*. Peirce termed our ability to selectively generate hypotheses out of this infinite space *abduction* and attributed this ability to an "instinct, resembling the instincts of the animals in its so far surpassing the general powers of our reason and for its directing us as if we were in possession of facts that are entirely beyond the reach of our senses" (par. 173).

In this paper, we investigate one possible type of cue that may partly underlie this abductive ability—cues from *event structure*. In the following sections, we first describe the features of causal relationships that are most critical to our goals of predicting and controlling events. Next, we introduce the idea of event hierarchies and describe how knowledge of event structure can be used productively to rule out hypotheses that are unlikely to possess these critical features of causation. Finally, we motivate our specific predictions and preview our empirical strategy.

Features of Causation

What makes an event or variable a good candidate for a cause? That is, what associations of cause and effect are most likely to pick out true and useful causal generalizations about the world? This normative issue has been of interest not only in philosophy of science (Salmon, 1984; Strevens, 2008; Woodward, 2005) and statistics (Pearl, 2000; Spirtes, Glymour, & Scheines, 1993), but also to many practitioners of natural and social sciences, such as epidemiology (Hill, 1965), genetics (Kendler, 2005), and economics (Hoover, 2001). Two generalizations to emerge from this discussion are that causal relationships are more useful to the extent that they are *insensitive to background conditions* and that they *pick out specific associations* (Woodward, 2010).

Insensitivity (also known as *invariance* or *robustness*) is a critical feature of useful causal generalizations (Lewis, 1973; Woodward, 2006). A causal relationship between *C* and *E* is *insensitive* to the extent that changes to the actual background conditions *B* do not disrupt the sufficiency of *C* to bring about *E*. That is, the nature of the causal relationship between *C* and *E* depends not only on facts about the truth of the counterfactual "given the actual background conditions *B*, if *C* occurs, then *E* would occur," but also on the truth of a range of counterfactuals of the form "given *different* background conditions *B**, if *C* occurs, then *E* would occur." To the extent that the latter sorts of counterfactuals are true, the relationship between *C* and *E* is relatively insensitive.

For example, suppose Suzy is throwing a stone at a bottle (Woodward, 2006). If Suzy throws the stone (*C*), then the bottle will fall over (*E*). The dependence relationship between *C* and *E* ("if Suzy throws the stone, then the bottle will fall over") is relatively insensitive because it is likely to hold up under changes to the background conditions, such as a strong gust of wind. In contrast, imagine Suzy was instead throwing a paper airplane at the bottle. This relationship is

relatively sensitive because minor changes to the background conditions such as a gust of wind are likely to disrupt the contingency between *C* and *E*.

Specificity is also widely thought to be a desirable feature of a causal relationship (Campbell, 2008; Kendler, 2005; Lewis, 2000; Woodward, 2010). A causal relationship is *specific* to the extent that changes to *C* result in precise and systematic changes to *E* (i.e., *E* has relatively few alternative causes) and to the extent that changes to *C* result in minimal changes to other variables (i.e., *C* has relatively few additional effects). In the extreme case, *C* and *E* might stand in a one-to-one relationship, such that *C* is the only cause of *E* and *E* is the only effect of *C*. For example, Huntington's disease has a very specific relationship with its genetic underpinnings, because it is controlled by a single gene. In contrast, height has a much less specific relationship with its genetic underpinnings because, despite being highly heritable, height is controlled by many genes.

Insensitivity and specificity are desirable features in a causal candidate because causal relationships with these properties are more conducive to precise and accurate inference. Given our uncertainty about the environment, insensitive causal relationships that are more robust in the face of environmental variation are more useful for predicting effects from causes, since the cause and effect are linked in a wider variety of circumstances. Similarly, causes that stand in a one-to-one relationship with their effects yield superior inferential power relative to many-to-one ratios, which may allow for impressive cause-to-effect prediction, but are ambiguous with respect to effect-to-cause diagnosis.

Further, causes with these properties are more useful as *control variables* for bringing about desired effects (Campbell, 2008; Lombrozo, 2010). Insensitive causes have reliable relationships with their effects under more diverse conditions, making them more widely

applicable strategies for controlling the environment. If Suzy wished to knock over the bottle, she is better off using the stone than the paper airplane because its control over the bottle's position is less sensitive to background conditions. Likewise, specific causes that stand in a one-to-one relationship are better control variables relative to those with one-to-many ratios, since a cause with many effects cannot be used to manipulate its effect without having many (potentially undesirable) consequences. The best interventions target their effect powerfully, but with minimal 'collateral damage'. Huntington's disease would likely be a far easier outcome to control through genetic therapies than height, because Huntington's disease stands in a more specific relationship with its genetic causes.

Since insensitive and specific causal relationships lead to improved abilities for inference and intervention, heuristics that rule out causal relationships lacking these features would potentially be able to narrow the space of candidate causes without much risk of ruling out useful causal generalizations. Next, we suggest a set of *event structure* heuristics that can pick out generalizations with these properties, and which can potentially be deployed in real time.

Event Structure

Hierarchical structure plays a broad role in our mental lives, influencing our preferences, predictions, and concepts of others (Trope & Liberman, 2003, 2010). Part-whole structures are especially critical for our perception of events (Baldwin, Baird, Saylor, & Clark, 2001; Newtson, 1973), and people are able to quickly and automatically segment events at relatively coarse and fine levels of granularity (Newtson, Hairfield, Bloomingdale, & Cutino, 1987; Zacks & Tversky, 2001). For instance, a social encounter may be perceived as distinct vocal utterances by each party at one level (i.e., as low-level events), while at the same time those utterances are grouped into conversational phases, such as an argument, an apology, gossiping, and so forth (i.e., as

higher-level events). This grouping of low-level events into higher-level events results in clusters of low-level events, and the higher-level events in turn can be grouped into yet higher-level events (see Figure 1). These discrete and hierarchical event representations are thought to play key roles in a variety of cognitive processes, including action planning (Newell & Simon, 1972), narrative comprehension (Zacks, Speer, & Reynolds, 2009), memory (Abbott, Black, & Smith, 1985), and communication (Zacks, Tversky, & Iyer, 2001).

This part-whole structure of events also contains information that could help restrict the hypothesis space in causal inference. Intuitively, events at the same hierarchical level are likelier to have a one-to-one correspondence with one another. Thus, low-level features (e.g., specific utterances) are good candidates for the causes of other utterances, and higher-level events within the conversation (e.g., a group of utterances that constitute an argument) are good candidates for the causes of other high-level events (e.g., the parties making up with one another). Parsing of events may also be useful for identifying portions of the event stream that are temporally close to the effect. Temporally proximal causes are likelier to reliably co-occur with the effect under different background conditions, since their pairing on this particular occasion is less likely to be a coincidence. Yet, what constitutes a temporally 'close' event is relative to the hierarchical level. The cause of a geological epoch may be another high-level event in the distant past, whereas the cause of a metabolic reaction may be another very recent low-level event. Thus, event structure may be useful in determining how widely to search for temporally proximal candidates. We can make these ideas more precise by identifying two ways that event structure can be used, which we term *level-matching* and *boundary-blocking*.

The *level-matching* heuristic assumes that a cause and its effect will tend to be at the same level of the event hierarchy. To return to our earlier example, it seems implausible to

identify the cause of Diane's lateness today as her car's spark plug firing late (too low in the hierarchy) or as her ongoing struggle with her car (too high in the hierarchy); instead, it seems more intuitive to identify her lateness as caused by her car's failure to start this morning. Level-matching helps to identify specific causal relationships because causes at the same hierarchical level are the most likely to stand in a one-to-one relationship to each other. That is, an event at a much lower level in the event hierarchy (like one of her car's spark plugs firing too late) is unlikely to be a good causal candidate for her lateness because it is too *fine* of a predictor—presumably, the spark plug fired late many times that morning, so why should we identify this particular firing as the cause? Conversely, an event at a much higher level in the hierarchy (like Diane's ongoing car problems) is a poor causal candidate for her lateness because it is too *coarse* of a predictor—if she is having an ongoing car struggle, why did it cause her to be late on this instance and not many others? Only an event at a similar hierarchical level to the effect will have the right 'resonance' to be a reliable predictor of the effect (see Lien & Cheng, 2000, for related arguments concerning taxonomies of categorization in causal learning).

The *boundary-blocking* heuristic assumes that a cause and effect (at the same hierarchical level) tend to both be parts of the same superordinate event; or, conversely, that two events at the same hierarchical level are less likely to be causally related if there is a higher-level event boundary between them. For example, Diane's being late and her alarm malfunctioning could be conceived as parts of a single superordinate episode (her Monday morning mishap), while events that happened earlier (e.g., her playing a game of bridge the previous night) would not share that superordinate, and hence would be poorer candidates for causes. Boundary-blocking helps to identify relatively insensitive relationships. Because an event *a1* will covary with its superordinate *A*, *a1* will tend to have its superordinate *A* as a background condition whenever it

occurs. Since other parts of *A* such as *a2* will also covary with *A*, *a2* will tend to co-occur with *a1* in a way that is relatively robust to changes in background conditions. In contrast, events belonging to other superordinates will not tend to have this insensitive relationship with *a1* because they are less likely to covary with *A*. That is, it is just a coincidence that Diane played bridge the night before her car mishap, and had circumstances been slightly different (e.g., had her bridge club met on Wednesdays rather than Sundays), these events might not have co-occurred at all. In contrast, her alarm going off and her being late to work are more likely to occur in the same superordinate episode, and this relationship would more likely be robust to changes in background conditions. Although related to temporal proximity, boundary-blocking is the more general cue—boundary-blocking will indeed tend to identify temporally proximal events as plausible candidates, but provides a non-arbitrary criterion for *how* proximal these events should be and explains why temporal proximity is useful.

In addition to helping focus on inductively useful regions of the hypothesis space, event structure cues also have desirable processing properties. Event segmentation appears to be at least partially automatic with event boundaries affecting the flow of information into and out of working memory (Zacks et al., 2009). This automatically processed information can therefore be made available at little additional computational cost for other purposes, such as causal inference. In addition, segmentation processes take advantage of both bottom-up cues from perceptual input and on top-down cues from the conceptual system (Baldwin, Andersson, Saffran, & Meyer, 2008; Buchsbaum, Canini, & Griffiths, 2011; Buchsbaum, Griffiths, Gopnik, & Baldwin, 2009; Zacks, 2004). Given these bottom-up influences, it is possible to gain inductive narrowing benefits from event structure even in highly novel situations about which we have little prior conceptual knowledge. Finally, event-parsing abilities are present in infancy

(Baldwin, Baird, Saylor, & Clark, 2001), suggesting that event structure cues could potentially play a role in early-emerging causal reasoning abilities and may form a robust foundation for such reasoning throughout the lifespan (e.g., Gweon & Schulz, 2011; Sobel & Kirkham, 2006).

In short, two broad structural properties of event sequences—levels that are potentially alignable and boundaries between successive large scale events—may provide a unique and powerful way of constraining inferences about cause-effect relationships. Because event perception is largely automatic, these benefits may often occur without explicit encoding of event structures and may impose modest cognitive loads, and because event perception is early-emerging, these cues are potentially within the grasp of quite young individuals. Since these structural effects of events largely transcend content, we expect that they will hold across a wide range of domains ranging from highly artificial and unfamiliar domains where content knowledge is essentially foreclosed, to more familiar domains in both the inanimate physical world and the social and interpersonal worlds. Given that a major proportion of our explicit causal inferences are about the traits, beliefs, and desires of others (e.g., Apperly, 2010; Uleman, Saribay, & Gonzalez, 2008) in meaning-laden contexts, the question arises as to whether aspects of event structures could still have an influence in interpersonal contexts. We argue here that they do and that indeed they may be especially important in the social realm where Peirce's problem may loom the largest of all.

The Current Experiments

In four experiments, we examine whether people use these proposed event structure cues. We tested the level-matching and boundary-blocking effects in unfamiliar event hierarchies in physical and social domains, to show that these event structure cues are applicable in a domain-general manner and in the absence of prior domain knowledge. We focused on cases where the

event structure was verbally described to the participant, to maximize experimental control over the particular event structure inferred by participants. Thus, participants read brief passages describing an event structure that was explicitly individuated for them (e.g., describing a chemical reaction composed of several sub-reactions or describing the rituals of an unfamiliar group of people). Consequently, these studies do not measure participants' causal inferences from 'experienced' events, but rather their causal inferences from 'described' events. However, the causal inferences produced by these experimental formats are generally consistent (e.g., Griffiths & Tenenbaum, 2005; Wasserman, 1990).

In Experiment 1, we tested the level-matching and boundary-blocking effects in diagnostic (effect-to-cause) reasoning. In addition to investigating the level-matching and boundary-blocking phenomena, we also examined whether information about the temporal order of the episodes increased participants' use of these strategies, and whether information about human agents decreased the use of these strategies. In Experiment 2, we extended these findings by asking participants to perform the reverse task—making inferences about the likely *effects* of events in the hierarchy.

In Experiments 3, we sought to replicate these effects using a different methodology, and to test an alternative account of level-matching in terms of *construal level* effects, which are especially prominent in the interpersonal sphere (Trope & Liberman, 2010). People often conceptualize events as psychologically distant or proximal on various dimensions, and these dimensions are interrelated. Thus, events described using more abstract language or at more superordinate levels are thought to be related to other events at that same level. Although construal level has not been implicated in causal judgments in the specific way we are studying, this interrelationship between events conceptualized at the same level could potentially lead to a

level-matching effect in causal inference. We addressed this possibility in Experiment 3 by looking for a case where level-matching would be predicted by a construal level account but not by our own, namely when two high-level events are not united by a common superordinate.

In Experiment 4, we sought to rule out a second alternative interpretation in terms of *similarity*. In the absence of other information, adults and children tend to look for causes that are similar to their effects in causal attribution (e.g., Einhorn & Hogarth, 1986; Shultz & Ravinsky, 1977; see White, 2009 for a review). A similarity-based account could potentially explain both the level-matching effect (since events at the same level are similar in terms of hierarchical level) and the boundary-blocking effect (since events that are part of the same superordinate are likely to share properties with each other). In Experiment 4, we addressed whether similarity can fully explain these effects by contrasting level-matching effects in partonomic structures (wherein subordinates are *parts* of their superordinates as in event hierarchies) and taxonomic structures (wherein subordinates are *kinds* of their superordinates and where similarity relationships might be expected to be especially potent).

Throughout these experiments, we anticipated that level-matching and boundary-blocking would be used to guide causal inferences, and that these effects would be modulated by event structure in the ways predicted by our event structure framework. In the General Discussion, we turn to the implications of these results for other issues related to causal and social reasoning more broadly.

Experiment 1

In Experiment 1, participants read paragraphs of information designed to capture the event hierarchy depicted in Figure 2, using three different cover stories—a chemistry story, a computer story, and a machine story. Unfamiliar events were used to remove any effects of prior

event-specific causal knowledge. In the primary condition of interest (the *Order* condition), participants read about a series of events that occurred in a fixed order. For example, the chemistry vignette was formatted as follows in that condition:

> Two chemicals were combined, which led to a series of chemical reactions. These occurred in the order given below.
>
> The first reaction was [*A*]. This consisted of three sub-reactions: first [*a1*], then [*a2*], then [*a3*]. The next reaction was [*B*]. This consisted of three sub-reactions: first [*b1*], then [*b2*], then [*b3*]. The next reaction was [*C*]. This consisted of three sub-reactions: first [*c1*], then [*c2*], then [*c3*].

In the stimuli seen by participants, the abbreviations (*A*, *a1*, *a2*, etc.) were replaced with names of real but likely unfamiliar chemical reactions (e.g., carbonylation, trimerisation) in one of two pseudorandom orders. The upper-case letters (*A*, *B*, *C*) represent the high-level or superordinate events, and their parts are represented by lower-case letters (e.g., the parts of *A* are *a1*, *a2*, and *a3*). The computer cover story differed in describing computer routines and sub-routines, and the machine story in describing machine processes and their sub-processes.

To test for level-matching, participants were asked to rate the likelihood that each of the other distinct events caused *C*, *c2*, and *c1*. Judgments about the high-level event *C* tested the level-matching principle for high-level effects: Our framework predicts that the high-level events (*A* and *B*) should be rated as better causes than their lower-level parts (*a1*, *a2*, *a3*, *b1*, *b2*, *b3*). Judgments about the low-level event *c2* tested the level-matching principle for low-level effects: Our framework predicts that in the Order condition, *c1* should be rated as the best cause (because *c1* is at the same level but occurred prior to *c2*), and in the No Order condition, *c1* and *c3* should be rated as the best causes (because *c1* and *c3* are both at the same level but it is unknown whether they happen before or after *c2* when order information is omitted). Finally, judgments about the low-level event *c1* tested the boundary-blocking effect: Low-level causes in different

event clusters (*a1*, *a2*, *a3*, *b1*, *b2*, *b3*) should be rated no higher than their high-level superordinates (*A* and *B*).

We additionally examined two possible boundary conditions on the level-matching and boundary-blocking effects. First, we varied whether the events were said to occur in the order in which they were stated (between the *Order* and *No Order* conditions). In the *No Order* condition, the chemistry vignette was formatted as follows:

> Two chemicals were combined, which led to a series of chemical reactions. These occurred in an unknown order.
>
> One of the reactions was [*A*]. This consisted of three sub-reactions: [*a1*], [*a2*], and [*a3*]. Another reaction was [*B*]. This consisted of three sub-reactions: [*b1*], [*b2*], and [*b3*]. Another reaction was [*C*]. This consisted of three sub-reactions: [*c1*], [*c2*], and [*c3*].

One possibility is that people rely on hierarchical information more when order information is omitted, because order information is itself an important cue to causal structure (e.g., Lagnado & Sloman, 2006). On the other hand, providing order information may encourage use of other temporal strategies such as the hierarchical cues we are examining. Thus, we asked whether the level-matching and boundary-blocking effects would be more, less, or equally strong given order information.

Second, we varied whether the events were said to occur spontaneously or with a human agent intervening at each event (between the *Order* and the *Order/Agent* conditions). In the *Order/Agent* condition, the chemistry vignette was formatted as follows:

> Fred, a chemist, conducted a series of chemical reactions. Fred conducted these reactions in the order given below.
>
> The first reaction Fred conducted was [*A*]. This consisted of his conducting three sub-reactions: first [*a1*], then [*a2*], then [*a3*]. The next reaction Fred conducted was [*B*]. This consisted of his conducting three sub-reactions: first [*b1*], then [*b2*], then [*b3*]. The next reaction Fred conducted was [*C*]. This consisted of his conducting three sub-reactions: first [*c1*], then [*c2*], then [*c3*].

Introducing an agent to each step might attenuate causal judgments for at least two reasons. First, people generally prefer to assign causal responsibility to human agents over non-human causes (Alicke, 1992; Hart & Honoré, 1985; Hilton, McClure, & Sutton, 2010; Lagnado & Channon, 2008). Because of this tendency, participants may tend to attribute causality of each reaction primarily to the agent and only secondarily to the other chemical reactions—a kind of discounting effect (Kelley, 1973) that could attenuate the overall ratings of each cause. Second, given that a human agent intervenes at each step, participants may see the reactions themselves as preconditions that *enable* rather than *cause* each reaction (Goldvarg & Johnson-Laird, 2001; Sloman, Barbey, & Hotaling, 2009; Wolff, 2007). Under these conditions where social causation (by the agent) might predominate over physical causation (by the other reactions), it is unclear whether hierarchical constraints would still be used.

Methods

Sixty-one participants were recruited from Amazon Mechanical Turk, and 19 participants were excluded from data analysis because they failed at least one of two manipulation check procedures (described below). However, in this and subsequent experiments, reported results are qualitatively the same if all participants are included in the analysis, except as indicated in footnotes.

Each participant answered questions about three vignettes, in a random order. The three conditions (*Order*, *No Order*, and *Order/Agent*) were assigned to the three vignette cover stories (chemical reactions, computer routines, or machine processes) using a Latin square. For each vignette, participants first read a brief passage that introduced the hierarchical structure. For example, the introduction for the chemistry vignette read: “On the next page, you will read about a series of chemical reactions that occurred. Some chemical reactions are made up of sub-

reactions. For example, if chemicals A and B react together to make chemical C, the overall reaction might be called W, but W might involve several distinct sub-reactions (e.g., x, y, z) in which various chemical intermediates are produced." Next, participants read the vignette (see above for examples). On the bottom of this page, participants completed a manipulation check task, in which they were presented with 10 pairs of events, and asked "Based on the information you read above, please identify the pairs below for which the second item is a sub-reaction of the first item." Parallel instructions were given for the computer and machine vignettes. There were thus a total of 30 manipulation check items across the three vignettes, and participants were excluded from data analysis if they answered more than 20% of these items incorrectly.

On the next page, participants completed the causal inference task. The text of the vignette was included at the top of this page to reduce the memory load of the task. For each vignette, they rated the likelihood that each of the other events caused the events *C*, *c1*, and *c2* (see Figure 2) on three separate screens, in a random order. On each screen, they were asked to "rate the extent to which you think each of the following reactions caused [*X*] to occur," where *X* was replaced with the *C*, *c1*, or *c2* event on each screen. These ratings were completed on a 9-point scale (1: "definitely did not cause"; 5: "unsure"; 9: "definitely did cause"). The candidates to be rated included all events that were logically distinct from the target event. That is, when rating the causes of *C*, participants were asked about *A*, *a1*, *a2*, *a3*, *B*, *b1*, *b2*, and *b3*; when rating the causes of *c1*, participants were asked about *A*, *a1*, *a2*, *a3*, *B*, *b1*, *b2*, *b3*, *c2*, and *c3*; and when rating the causes of *c2*, participants were asked about *A*, *a1*, *a2*, *a3*, *B*, *b1*, *b2*, *b3*, *c1*, and *c3.* The events to be rated as causes were always listed in the same order as they were listed in the vignette (e.g., for the question about *C*, the events were rated in the order *A*, *a1*, *a2*, *a3*, *B*, *b1*, *b2*, *b3*).

After completing all three vignettes, participants answered five additional multiple-choice manipulation check questions to ensure that they had attended to which vignettes had included information about temporal order and agency. Participants who answered three or more of these questions incorrectly were excluded from analysis.

Results

High-high level-matching (judgments about *C*). To test for high-high level-matching, we examined responses to the question about the causes of the high-level event *C*. As shown in Figure 3, high-high level-matching occurred: the high-level predecessors (*A* and *B*) were rated as better causes of *C* than were their subordinate events (*a1*, *a2*, *a3*, *b1*, *b2*, *b3*).

For statistical tests, we calculated each participant's mean rating for the high-level predecessors (*A* and *B*) and for the low-level predecessors (*a1*, *a2*, *a3*, *b1*, *b2*, *b3*), and entered these scores into a repeated measures ANOVA, with event type (high, low) and condition (Order, No Order, Order/Agent) as within-subjects factors.[1] A main effect was obtained for event type, $F(1,41) = 11.51$, $p = .002$, $\eta_p^2 = .22$, with high-level predecessors rated as better causes than low-level predecessors ($M = 5.22$, $SD = 2.24$ vs. $M = 4.80$, $SD = 2.15$). This is the predicted level-matching effect.

A main effect was also obtained for condition, $F(2,82) = 14.49$, $p < .001$, $\eta_p^2 = .26$, because causal ratings were highest in the Order condition ($M = 6.14$, $SD = 2.75$), followed by the Order/Agent condition ($M = 4.98$, $SD = 2.99$), followed by the No Order condition ($M = 3.90$, $SD = 2.18$). The lower rating in the Order/Agent condition compared to the Order condition makes sense in light of previous work on agents (e.g., Lagnado & Channon, 2008), which would

[1]In some cases, the sphericity assumption was violated for repeated measures ANOVAs reported in Experiments 1 and 2. Application of Greenhouse-Geisser corrections did not affect the significance level of any result, however, so we report the uncorrected tests for ease of exposition.

suggest that people might attribute causality to the *agent* rather than to the *physical cause* when an agent is available as an explanation. The lower rating in the No Order condition compared to the Order condition most likely resulted from a decrease in confidence when the causal judgments were completed without the benefit of temporal order information, which is an important causal cue (Lagnado & Sloman, 2006). However, there was no interaction between event type and condition, $F(2,82) = 0.74$, $p = .48$, $\eta_p^2 = .02$. Thus, even when social information trumps physical information (as in the Order/Agent condition) or when confidence is undermined by a lack of temporal information (as in the No Order condition), hierarchical cues can still be used to prune the hypothesis space.

Low-low level-matching (judgments about *c2*). To test for low-low level-matching, we examined responses to the question about the causes of *c2*. As shown in Figure 4, low-low level-matching occurred: *c2*'s low-level predecessor *c1* was rated much higher than any other event. Boundary-blocking also occurred, in that low-level events in other clusters (i.e., *a1*, *a2*, *a3*, *b1*, *b2*, *b3*) were rated no higher than their superordinates (*A* and *B*).

For statistical tests, we calculated each participant's mean rating for the high-level predecessors (*A* and *B*), for the low-level predecessors in other clusters (*a1*, *a2*, *a3*, *b1*, *b2*, *b3*), for the low-level predecessor from the same cluster (*c1*), and for the low-level successor from the same cluster (*c3*). These scores were entered into a repeated measures ANOVA, with event type (high-other, low-other, low-same-predecessor, low-same-successor) and condition (Order, No Order, Order/Agent) as factors. Main effects were obtained for event type, $F(3,123) = 41.18$, $p <$

.001, $\eta_p^2 = .50$, and for condition, $F(2,82) = 4.80$, $p = .011$, $\eta_p^2 = .11$, as well as a significant interaction between event type and condition, $F(6,246) = 11.43$, $p < .001$, $\eta_p^2 = .22$.[2]

The main effect of event type occurred because the low-level predecessor from the same cluster was rated highest (*c1*; M = 6.31, *SD* = 2.12), followed by the high-level predecessors in other clusters (*A* and *B*; M = 5.11, *SD* = 2.29), followed by the low-level predecessors in other clusters (*a1*, *a2*, *a3*, *b1*, *b2*, *b3*; *M* = 4.75, *SD* = 2.14), followed by the low-level successor from the same cluster (*c3*; *M* = 2.25, *SD* = 1.46). That is, the ratings were higher for *c1* than for any other event, consistent with low-low matching. Moreover, low-level events from other clusters were not rated any higher than their high-level superordinates (in fact, they were rated lower), consistent with boundary-blocking.

As shown in Figure 4, the interaction between event type and condition was primarily driven by ratings of *c3*. While *c3* was rated lower than *c1* in all three conditions, this effect was the strongest in the Order and Order/Agent conditions, in which temporal order was available, $t(41) = 9.92$, $p < .001$, $d = 1.53$ and $t(41) = 8.88$, $p < .001$, $d = 1.37$, respectively, and was relatively weaker in the No Order condition, $t(41) = 4.02$, $p < .001$, $d = 0.62$. It is unsurprising that this effect was weaker in the No Order condition, because either *c1* or *c3* could have occurred before or after *c2*. The fact that *c1* was still rated a significantly better cause than *c3* suggests that participants could not completely override the temporal cue of *c1* being listed before *c3*.

[2] For low-low level-matching, the main effect of condition is reduced to marginal significance if all participants are included in the analysis, both for Experiment 1 [$F(2,120) = 2.27$, $p = .107$, $\eta_p^2 = .04$] and Experiment 2 [$F(2,118) = 2.34$, $p = .101$, $\eta_p^2 = .04$]. However, because this analysis includes participants who failed the manipulation checks and may not have understood the differences among the three conditions, the more exclusive analysis is more appropriate here.

Boundary-blocking (judgments about *c1*). To test for boundary-blocking, we examined responses to the question about the causes of *c1*. As shown in Figure 5, boundary-blocking occurred: unlike in the case of rating causes for *c2* (cf. Figure 4), the event's immediate low-level predecessor, *b3*, was not the highest-rated cause. Indeed, *c1*'s high-level predecessors *A* and *B* tended to be rated as more likely causes than their subordinate events.

For statistical tests, we calculated each participant's mean rating for the high-level predecessors (*A* and *B*), for the low-level predecessors in other clusters (*a1*, *a2*, *a3*, *b1*, *b2*, *b3*), and for the low-level successors from the same cluster (*c2* and *c3*). These scores were entered into a repeated measures ANOVA, with event type (high-other, low-other, low-same) and condition (Order, No Order, Order/Agent) as factors. A main effect was obtained for event type, $F(2,82) = 29.11$, $p < .001$, $\eta_p^2 = .42$, but not for condition, $F(2,82) = 1.88$, $p = .16$, $\eta_p^2 = .04$; however, a significant interaction occurred between event type and condition, $F(4,164) = 16.86$, $p < .001$, $\eta_p^2 = .29$.

The main effect of event type occurred because the high-level predecessors from other clusters were rated highest (*A* and *B*; $M = 5.05$, $SD = 2.39$), followed by the low-level predecessors from other clusters (*a1*, *a2*, *a3*, *b1*, *b2*, *b3*; $M = 4.73$, $SD = 2.17$), followed by the low-level successors from the same cluster (*c2* and *c3*; $M = 2.36$, $SD = 1.70$). This is not the pattern to be expected if level-matching occurred across event boundaries; if that were the case, the low-level predecessors would be rated highest, but in fact they were rated lower than their high-level superordinates.

The interaction between event type and condition occurred primarily due to the ratings of the low-level successors *c2* and *c3*. In the Order condition, the low-level successors were rated lower than the low-level predecessors ($M = 1.67$, $SD = 1.90$ vs. $M = 5.50$, $SD = 2.88$; $t(41) =$

−6.60, $p < .001$, $d = -1.02$), as well as in the Order/Agent condition ($M = 1.94$, $SD = 2.01$ vs. $M = 4.95$, $SD = 2.88$; $t(41) = -4.90$, $p < .001$, $d = -0.76$), while ratings of these event types did not differ in the No Order condition ($M = 3.46$, $SD = 2.65$ vs. $M = 3.74$, $SD = 2.37$; $t(41) = -0.50$, $p = .62$, $d = -0.08$). This effect can be explained as follows: In the conditions with order information, participants could triage *c2* and *c3* as possible causes, because they occurred after *c1* and were thus even less likely to be causes than low-level predecessors from previous clusters. In the No Order condition, however, participants are pushed in two directions—on the one hand, *c2* and *c3* might have occurred after *c1*, in which case they are much worse causal candidates than the low-level predecessors; on the other hand, *c2* and *c3* might have occurred before *c1*, in which case they are much better candidates than the low-level predecessors, because they are in the same cluster. The balance of these opposing forces would lead participants to rate the low-level predecessors and successors similarly in the No Order condition.

Discussion

Overall, these results are as predicted by the proposed framework for the use of event structure cues in constraining Peirce's problem of hypothesis generation. Level-matching occurred robustly, with participants matching high-level causes (*A* and *B*) to high-level effects (*C*), and low-level causes (*c1*) to low-level effects (*c2*). Boundary-blocking also occurred robustly, with causes from other clusters (*a1*, *a2*, *a3*, *b1*, *b2*, *b3*) rated no higher than their superordinates for low-level effects (*c1* and *c2*).

Hierarchical cues were used to prune the hypothesis space even when confidence in causal judgments was undermined by omitting temporal order information (in the No Order condition) or by introducing an agent intervening at each step (in the Order/Agent condition),

providing a social cause that trumped the physical cause (Lagnado & Channon, 2008). Both manipulations resulted in lower overall causal judgments across all candidate causes, but the *relative* plausibility of the candidates was influenced by the same hierarchical cues participants used in the less ambiguous Order condition.

One aspect of these results worth noting is that judgments were frequently below the midpoint on our scales (see Figures 3–5), particularly in the Order/Agent and No Order conditions. This is entirely consistent with the use we are suggesting for event structure cues and with our experimental set-up. Participants are faced with a difficult task of assessing the relative plausibility of many different causes, and the evidence is highly ambiguous with respect to which is the true cause of each effect. Indeed, event structure and temporal order were the only cues available for making these judgments. With the possible exception of low-low level-matching (i.e., the strong preference of *c1* as a cause of *c2* in Figure 4), event structure cues appear to be used principally to prune the hypothesis space rather than to arrive at certain answers. To make confident causal judgments, participants would likely need more definitive cues such as mechanism knowledge (Ahn, Kalish, Medin, & Gelman, 1995) or statistical evidence (Cheng, 1997). Nonetheless, participants used event structure cues in a robust manner to modulate their judgments about plausibility in ways that could privilege one or two candidates, making the computational challenge of assessing those privileged candidates far more tractable.

Experiment 2

Diagnostic inference (reasoning from effect to cause) and predictive inference (reasoning from cause to effect) are not mirror images of one another (e.g., Pearl, 1988; Waldmann & Holyoak, 1992). In particular, multiple causes compete with one another as potential

explanations for an effect, whereas multiple effects do not compete with one another as potential consequences of a cause (Pearl, 1988). For example, in diagnosing the cause of Diane's lateness to work, we might imagine several possible causes (her alarm not going off, her car failing to start), but upon learning that one of these causes is in operation, we would think the other cause is unlikely—the *explaining away* or *discounting* effect (Kelley, 1973). However, in predicting the effects of Diane's lateness (a reprimand from her boss, a missed phone call), learning that one effect has occurred does not diminish our confidence in the others (Pearl, 1988). For this reason, we may be likely to focus our attention on identifying a few plausible causes in diagnostic reasoning, but willing to spread out our inferences about potential effects in predictive reasoning over a larger number of candidates. Indeed, this tendency is particularly pronounced for physical causal systems such as those used in Experiment 1. People identify a small number of causes for physical events but a larger number of effects, whereas for social events estimates of the number of causes and effects are more symmetric (Strickland, Silver, & Keil, 2014).

Given this reduced pressure in predictive inference to identify unique effects of a cause, a strategy such as level-matching might be less likely to occur in prediction. Replicating level-matching in a predictive inference task would thus be an especially strong test of event structure cues. Experiment 2 therefore used the same vignettes as Experiment 1 but reversed the task, asking participants to instead infer the effects of several potential causes.

Methods

Sixty participants were recruited from Amazon Mechanical Turk in exchange for a small payment; 19 participants were excluded from data analysis because they failed to meet the same exclusion criteria used in Experiment 1.

Participants read exactly the same vignettes as in Experiment 1. The design and procedure were identical to Experiment 1, except that rather than rating the likelihood that the other events were causes of *C*, *c2*, and *c1*, participants instead rated the likelihood that the other events were *caused by A*, *a2*, and *a3*. For example, for the chemistry vignette, they were asked to "rate the extent to which you think [*X*] caused each of the following reactions to occur," where *X* was replaced with *A*, *a2*, or *a3*.

Results

High-high level-matching (judgments about *A*). To test for high-high level-matching, we examined responses to the question about the effects of *A*. As shown in Figure 6, high-high level-matching occurred: the high-level successors (*B* and *C*) were rated as more likely to be caused by *A* than were their subordinate events.

For statistical tests, we calculated each participant's mean rating for the high-level successors (*B* and *C*) and for the low-level successors (*b1*, *b2*, *b3*, *c1*, *c2*, *c3*), and entered these scores into a repeated measures ANOVA, with event type (high, low) and condition (Order, No Order, Order/Agent) as factors. This revealed a main effect of event type, $F(1,40) = 4.83$, $p = .034$, $\eta_p^2 = .11$, and a main effect of condition, $F(2,80) = 6.51$, $p = .002$, $\eta_p^2 = .14$, with no interaction between event type and condition, $F(2,80) = 0.47$, $p = .63$, $\eta_p^2 = .01$. This is similar to the pattern of results obtained in Experiment 1 for diagnostic inference: high-level effects were rated as more likely to be caused by *A* than low-level effects ($M = 4.72$, $SD = 1.99$ vs. $M = 4.32$, $SD = 2.00$), and ratings were highest in the Order condition ($M = 5.40$, $SD = 2.79$), followed by the Order/Agent condition ($M = 4.30$, $SD = 2.68$), followed by the No Order condition ($M = 3.87$, $SD = 1.96$).

Low-low level-matching (judgments about *a2*). To test for low-low level-matching, we examined responses to the question about the effects of *a2*. As shown in Figure 7, low-low level-matching occurred: *a2*'s low-level successor *a3* was rated much higher than any other event. Boundary-blocking also occurred, in that low-level events in other clusters (i.e., *b1*, *b2*, *b3*, *c1*, *c2*, *c3*) were rated no higher than their superordinates (*B* and *C*).

For statistical tests, we calculated each participant's mean rating for the high-level successors (*B* and *C*), for the low-level successors in other clusters (*b1*, *b2*, *b3*, *c1*, *c2*, *c3*), for the low-level predecessor from the same cluster (*a1*), and for the low-level successor from the same cluster (*a3*). These scores were entered into a repeated measures ANOVA, with event type (high-other, low-other, low-same-predecessor, low-same-successor) and condition (Order, No Order, Order/Agent) as factors. Main effects were obtained for event type, $F(3,120) = 40.31$, $p < .001$, $\eta_p^2 = .50$, and for condition, $F(2,80) = 3.79$, $p = .027$, $\eta_p^2 = .09$, which were qualified by an interaction between event type and condition, $F(6,240) = 22.32$, $p < .001$, $\eta_p^2 = .36$.

The main effect of event type occurred because the low-level successor from the same cluster was rated highest (*a3*; $M = 6.44$, $SD = 1.90$), followed by the high-level successors in other clusters (*B* and *C*; $M = 4.13$, $SD = 1.99$) and low-level successors in other clusters (*b1*, *b2*, *b3*, *c1*, *c2*, *c3*; $M = 4.03$, $SD = 1.95$), followed by the low-level predecessor from the same cluster (*a1*; $M = 2.63$, $SD = 1.64$). That is, the ratings were higher for *a3* than for any other event, consistent with low-low matching.

The interaction between event type and condition was primarily driven by ratings of *a1*. Although *a1* was rated lower than *a3* in all three conditions, this effect was the strongest in the Order and Order/Agent conditions, in which temporal order was available, $t(40) = -11.49$, $p < .001$, $d = -1.79$ and $t(40) = -9.07$, $p < .001$, $d = -1.42$, respectively, and was relatively weaker in

the No Order condition, $t(40) = -3.62$, $p = .001$, $d = -0.57$. As in Experiment 1, it makes sense that this difference should be smaller in the No Order condition, where *a1* and *a3* could have occurred either before or after *a2*.

Boundary-blocking (judgments about *a3*). To test for boundary-blocking, we examined responses to the question about the effects of *a3*. As shown in Figure 8, boundary-blocking occurred: the event's immediate low-level successor, *b1*, was not rated as more likely to be caused by *a3* than any other successor event.

For statistical tests, we calculated each participant's mean rating for the high-level successors (*B* and *C*), for the low-level successors in other clusters (*b1*, *b2*, *b3*, *c1*, *c2*, *c3*), and for the low-level predecessors from the same cluster (*a1* and *a2*). These scores were entered into a repeated measures ANOVA, with event type (high-other, low-other, low-same) and condition (Order, No Order, Order/Agent) as factors. A main effect was obtained for event type, $F(2,80) = 14.54$, $p < .001$, $\eta_p^2 = .27$, because the high-level successors from other clusters (*B* and *C*; $M = 4.27$, $SD = 2.08$) and low-level successors from other clusters (*b1*, *b2*, *b3*, *c1*, *c2*, *c3*; $M = 4.18$, $SD = 2.09$) were rated higher than the low-level predecessors from the same cluster (*a1* and *a2*; $M = 2.54$, $SD = 1.63$). This pattern is not what would be expected if level-matching had occurred across the event boundary—in that case, one would expect low-level successors to be rated higher than high-level successors, which was not observed.

The main effect of condition was only marginally significant, $F(2,80) = 2.59$, $p = .081$, $\eta_p^2 = .06$, but there was a significant interaction between event type and condition, $F(4,160) = 23.13$, $p < .001$, $\eta_p^2 = .37$. This occurred because low-level predecessors (*a1*, *a2*) were rated lower than low-level successors (*b1*, *b2*, *b3*, *c1*, *c2*, *c3*) in the Order and Order/Agent conditions, $t(40) = -5.73$, $p < .001$, $d = -0.90$ and $t(40) = -3.49$, $p < .001$, $d = -0.55$, respectively, but were

rated equally in the No Order condition, $t(40) = 0.91$, $p = .37$, $d = 0.14$. This pattern is the same as that found in Experiment 1, and most likely occurred because participants in the No Order condition were unsure whether *a1* and *a2* occurred before *a3* (in which they are less likely to be effects than the low-level successors from other clusters) or after *a3* (in which case they are more likely to be effects than the low-level successors from other clusters).

Discussion

These results fully replicate the findings of Experiment 1 in predictive (cause-to-effect) inference. Both matching effects and boundary-blocking occurred robustly, and regardless of whether order information or agent information were given. The main effects of condition from Experiment 1 were also replicated—participants were consistently less confident in predicting effects when order information was omitted, and when an agent was said to have initiated each step. Thus, event structure is used to constrain both diagnostic and predictive causal inference.

Experiment 3

Events that are distant in time tend to be conceptualized in relatively abstract ways, compared to events that are more proximal (Trope & Liberman, 2003). Further, the various dimensions of psychological distance (e.g., time, space, hypotheticality) are cognitively related, so that thinking of an event at a higher level can make it feel more temporally distant (Trope & Liberman, 2010). Construal level theory (CLT) can therefore provide an alternative explanation of the level-matching effect in Experiments 1 and 2—namely, that high-level events would be thought of as more psychologically distant, requiring more psychologically distant causes (in Experiment 1) or causing more psychologically distant effects (in Experiment 2). Although CLT's explanation for boundary-blocking is less clear, Experiment 3 sought to rule out the CLT

explanation for level-matching and to extend our previous results to a new event structure and new domain, using a simpler task.

To distinguish between our event structure account and the CLT account, we looked for an event structure where we would not predict level-matching, but where a CLT account would. In particular, our account would not predict high-high level-matching when two high-level events do not share a common superordinate event (see Figure 9). The structure depicted in Figure 9A is similar to those used in our previous experiments, in that those structures always included a unifying 'super-superordinate' event (e.g., a chemical reaction, a mechanical process). In Figure 9B, in contrast, the structure differs in failing to have a super-superordinate event α to unite *A* and *B* events. Because *A* and *B* are not related in the same event structure, we would not expect level-matching for this structure. However, in both structures, *A* and *B* are both situated at relatively high levels of their hierarchies, so the CLT account of level-matching would predict no difference in level-matching between these structures.

We instantiated these structures in a social domain—the practices of an unfamiliar group of people named the Favonians, who live in a distant country. This domain and simpler event structure allowed us to use more naturalistic wordings and less repetitive stimuli, in a domain that would be more familiar to participants. Extending the results of Experiments 1 and 2, which involved primarily physical causation (e.g., chemical reactions and machine operations), to social causation also helps to establish the generality of our findings. This is especially important in light of findings that people often think about social causation in a rather different way from physical causation—as more teleological rather than mechanistic (e.g., Lombrozo, 2010), more stochastic rather than deterministic (e.g., Johnson & Keil, 2014; Strickland, Silver, & Keil, 2014), and more rooted in counterfactual dependence than in transference of force (e.g.,

Lombrozo, 2010). Although Experiments 1 and 2 showed that people still use event structure cues when agents intervene on physical systems, these cues may no longer be used in a system where all causal links are socially construed. Alternatively, because the apparent universe of cause and effects may seem especially large in social systems, it may be that event structure cues are especially needed in such cases to help narrow down candidate causes and effects.

The key hypothesis was that high-high level-matching (preferring *A* over *a1* as a cause of *B*) would be stronger for the structure in Figure 9A than the structure in Figure 9B. To reduce the load of the task, rather than asking about all possible causes of the events in the hierarchy, we instead focused on individual cases where a Favonian had done three of the practices from the hierarchy (e.g., "Lee is a Favonian who [*B*]ed, [*A*]ed, and [*a1*]ed," where the abbreviations were replaced with novel names for these practices such as 'kwerp' and 'gorn'). To test high-high level-matching, participants were asked whether *A* or *a1* was a more likely cause of *B*. We would expect that for the hierarchy depicted in Figure 9A (used in Experiment 3A) that includes the super-superordinate, high-high level-matching would occur, so participants would think *A* is a more likely cause than *a1*. However, for the hierarchy depicted in Figure 9B (used in Experiment 3B), we would expect high-high level-matching to be weaker or null, since *A* and *B* are not embedded in the same structure. We also tested for low-low level-matching and boundary-blocking, to see whether these effects would generalize to this different task. We did not anticipate any differences in low-low level-matching or boundary-blocking between the structures in Figures 9A and 9B, since the relevant structural features (i.e., at the low and middle levels of the hierarchy) are the same between these structures.

Methods

We recruited 246 participants from Amazon Mechanical Turk to participate in Experiment 3, with 127 randomly assigned to Experiment 3A and 119 randomly assigned to Experiment 3B. Forty participants from Experiment 3A and 34 participants from Experiment 3B were excluded from data analysis because they incorrectly answered more than 20% of the check questions (see below).

In the *hierarchy induction* phase of Experiment 3A, participants learned about the event structure depicted in Figure 9A (letters in brackets refer to Figure 9A, but these were not available to participants):

> In a distant country, there is a group of people called the Favonians, with a unique culture involving practices such as *merking*, *zorbing*, *kwerping*, *wanning*, *gorning*, *lepping*, and *qualfing*.
>
> Sometimes some of these practices can trigger other practices, causing them to occur.
>
> Some of the Favonians' practices are steps involved in other practices. For example, in the United States, we have a practice of eating meals. Two of the steps involved in eating a meal are putting out the dishes and having dessert.
>
> Favonians sometimes kwerp [*a1*] and wann [*a2*], which are both steps involved in zorbing [*A*]. Favonians also sometimes lepp [*b1*] and qualf [*b2*], which are both steps involved in gorning [*B*].
>
> In addition, zorbing [*A*] and gorning [*B*] are both steps involved in merking [α].

Two pseudorandom orders were used to assign the blank verbs (e.g., *lepp* and *qualf*) to *A*, *a1*, *a2*, *B*, *b1*, and *b2*. On the same screen as this information, participants completed a series of 16 true/false check questions to ensure that they encoded the event structure (e.g., "Kwerping is a step involved in zorbing"). The hierarchy induction was identical for Experiment 3B, except the α event (*merking*) was not mentioned, and the last paragraph of the above instructions were deleted. Due to the simpler hierarchy, participants in Experiment 3B only competed 12 true/false check questions.

On the following screens, participants completed a total of three causal inference questions, concerning the causes of *B* (to test high-high level-matching), the causes of *b2* (to test low-low level-matching), and the causes of *b1* (to test boundary-blocking). These three questions were completed in a random order, and all the information the participants learned about event structure was included at the top of each screen to reduce memory demands.

For the *high-high level-matching* question, participants read the following:

> Lee is a Favonian who [*B*]ed, [*A*]ed, and [*a1*]ed.
>
> Lee's [*B*]ing was caused by either his [*A*]ing or his [*a1*]ing, but we aren't sure which. Which do you think is more likely to have caused Lee to [*B*]?

Responses were entered on a scale from 1 ("Definitely [*A*]ing") to 9 ("Definitely [*a1*]ing"). The order of mentioning *A* and *a1* in the question was counterbalanced, and the left/right counterbalancing of the response scale was adjusted to match this order. The *low-low level-matching* and *boundary-blocking* questions were formatted similarly, but concerned different individuals who completed different sets of practices. For the *low-low level-matching* question, participants read about Bruce, who did *b2*, *A*, and *b1*. Participants rated the likelihood that *b2* was caused by *A* or by *b1*. For the *boundary-blocking* question, participants read about Sarah, who did *b1*, *A*, and *a2*. Participants rated the likelihood that *b1* was caused by *A* or by *a2*. These questions were identical in Experiments 3A and 3B. Thus, the only difference between these experiments was in whether a super-superordinate event α was mentioned in the induction phase that subsumed *A* and *B*.

After completing the causal inference items, participants answered 16 additional true/false check questions (e.g., "Bruce is a Favonian") to ensure that they were attending to the task. Any participant who incorrectly answered more than 20% of the combined sets of check questions was excluded from data analysis.

Results and Discussion

When a 'super-superordinate' event (α) was specified, participants in Experiment 3A behaved consistently with those in our previous experiments, showing a high-high level-matching effect, a low-low level-matching effect, and a boundary-blocking effect. In contrast, when the α event was not given, participants in Experiment 3B no longer showed high-high level-matching. Yet, they showed low-low level-matching and boundary-blocking to the same degree as participants in Experiment 3A. These findings are depicted in Figure 10.[3]

Responses for each question were converted to a scale from -4 to 4, where negative scores corresponded to endorsements of the low-level cause, and positive scores corresponded to endorsements of the high-level cause, with 0 indicating no preference. To test high-high level-matching, we analyzed responses to the question about causes of *B*, anticipating that level-matching would lead participants to prefer the high-level event *A* over the low-level event *a2*. In Experiment 3A, where a super-superordinate event α was specified, participants' preference for *A* over *a2* ($M = 1.30$, $SD = 2.28$) led to significantly positive scores, $t(86) = 5.32$, $p < .001$, $d = 0.57$. In Experiment 3B, however, where α was not specified, participants' preference for *A* ($M = 0.47$, $SD = 2.27$), reached only marginal significance, $t(84) = 1.90$, $p = .061$, $d = 0.21$. This also led to a significant difference between Experiments 3A and 3B, $t(170) = 2.41$, $p = .017$, $d = 0.37$, because high-high level-matching was significantly stronger when an α event was given.

To test low-low level-matching, we analyzed responses to the question about causes of *b2*, anticipating that level-matching would lead participants to prefer the low-level event *b1* over

[3] For some dependent measures used in Experiments 3, Kolmogorov-Smirnov tests revealed non-normal distributions. All one-sample *t*-tests in Experiments 3 were repeated using non-parametric one-sample Wilcoxon tests, and all independent-samples *t*-tests in Experiment 3 repeated using non-parametric Mann-Whitney tests. In every case, the significance level was in agreement with the parametric test, so we report the more straightforward *t*-tests in the text.

the high-level event *A*. In both Experiments 3A and 3B, low-low level-matching occurred, with participants in both experiments significantly preferring *b1* over *A* (M = -1.30, SD = 2.50 and $t(86) = -4.83$, $p < .001$, $d = -0.52$ for Experiment 3A; M = -1.69, SD = 2.29 and $t(84) = -6.81$, $p < .001$, $d = -0.74$ for Experiment 3B). Moreover, this effect was equally strong in both experiments, $t(170) = 1.08$, $p = .28$, $d = 0.16$.

Finally, to test for boundary blocking, we analyzed responses to the question about causes of *b1*, anticipating that participants would no longer have a preference for a low-level cause (*a2*) over a high-level cause (*A*) because the low-level cause belongs to a different superordinate. As in Experiments 1 and 2, boundary-blocking actually led participants to have a slight preference for the *high*-level cause *A*, leading to marginally positive scores (M = 0.45, SD = 2.24 and $t(86) = 1.89$, $p = .063$, $d = 0.20$ for Experiment 3A; M = 0.47, SD = 1.98 and $t(84) = 2.17$, $p = .033$, $d = 0.24$ for Experiment 3B). This effect did not differ in strength between experiments, $t(170) = 0.04$, $p = .97$, $d < 0.01$.

These results count against a construal level account of the level-matching effect found in Experiments 1 and 2. If people causally matched high-level events to other high-level events because they were at the same level of abstractness or because they were seen as more temporally distant than the low-level events, they should have done so in both the structures used in Experiment 3. Instead, high-high level-matching occurred only for the structure depicted in Figure 9A, where a super-superordinate event unified the *A* and *B* events. However, the low-low level-matching and boundary-blocking effects were consistent across both structures, just as we would expect given that the low and middle levels of these hierarchies are identical. Finally, these results go beyond those in Experiments 1 and 2 by showing that event structure cues are used not just for making inferences about physical causal systems, but also for social systems.

Experiment 4

In our final experiment, we addressed one further alternative account of event structure cues—similarity matching. A classic idea in anthropology is the *principle of homeopathy*—that causes will resemble their effects (Frazer, 1959; Shweder, 1977). Both adults and children use similarity matching to make causal inferences, using a variety of dimensions of similarity including intensity, size, and even color (e.g., Einhorn & Hogarth, 1986; LeBoeuf & Norton, 2012; Rozin, Millman, & Nemeroff, 1986; Shultz & Ravinsky, 1977). Indeed, the visual system uses similarity of the size and speed of objects in judgments of perceptual causality (Michotte, 1963/1946), suggesting that resemblance-based causal reasoning may not only be culturally widespread or universal (Shweder, 1977) but in some cases even rooted in the deepest and earliest emerging routines of visual processing (see White, 2009 for a review).

To what extent can our current findings be explained in terms of similarity matching? Medin (1989; see also Goodman, 1955 and Medin, Goldstone, & Gentner, 1993) noted that similarity alone is too unconstrained a notion to do explanatory work without invoking further principles. In particular, similarity-based theories must specify which *dimensions* are used to compute similarity. In the case of event structure cues, two principles would be needed—a principle of similarity among levels (to explain level-matching) and a principle that subordinate nodes inherit features of their superordinates (to explain boundary-blocking). This second principle is necessary to explain how low-level events, despite being at the same level of analysis, are not matched when they belong to different superordinate events. If events are assumed to inherit properties from their superordinates, then low-level events will share more features in common when they belong to a common superordinate, so similarity-matching could yield stronger inferences when low-level events belong to the same superordinate.

In some sense, the similarity-based explanation accounts for hierarchy-based effects at a different level of analysis, and so does not necessarily compete with our own account. Further, the principles furnished to constrain similarity are ultimately principles about event structure, so a similarity-based account may actually require the very principles we are arguing for. Nonetheless, event structure cues would be especially powerful if they can actually *override* similarity in making causal attributions.

In Experiment 4, we pitted two different types of matching strategies against one another—a *partonomic* or *event structure* match on the one hand, against a *taxonomic* match on the other hand (see Tversky & Hemenway, 1984). In a *partonomic* hierarchy, lower-level nodes are *parts* of higher-level nodes (e.g., a tail is a *part* of a dog), whereas in a *taxonomic* hierarchy, lower-level nodes are *kinds* of higher-level nodes (e.g., a German Shepherd is a *kind* of dog). Not only objects, but events can also have partonomic or taxonomic structures (e.g., having dessert is a *part* of eating a meal, whereas breakfast is a *kind* of meal). The structures used in Experiments 1–3 were partonomies, because the lower-level event were sub-events of the higher-level events.

In Experiment 4, we asked whether the level-matching principle would be stronger for partonomic rather than taxonomic matches. Given that the logic of event structure cues depends on part-whole relationships between events (Campbell, 2008; Lewis, 2000; Woodward, 2006), and the computational advantage of event structure cues depends on the perceptual extraction of this structure from experience (e.g., Buchsbaum et al., 2009; Zacks, 2004), we would expect partonomic matches to be stronger than taxonomic matches for inferring causal structure. In contrast, taxonomic matches could well be stronger than partonomic matches in terms of similarity because they often entail large sets of common perceptual features (e.g., Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). Although we know of no studies that have

directly pitted partonomic versus taxonomic similarity against each other, people do rely strongly on taxonomy-based similarity cues, for example in inductive reasoning (e.g., Heit, 2000). If people preferred partonomic matches in causal judgments but taxonomic matches in similarity judgments, this would not only affirm that event structure principles are necessary to constrain a similarity-based account, but can even *override* similarity.

Methods

We recruited 120 participants from Amazon Mechanical Turk to participate in Experiment 4A and another 120 participants to recruit in Experiment 4B; 53 participants from Experiment 4A and 48 participants from Experiment 4B were excluded because they incorrectly answered more than 20% of the check questions.

Like Experiment 3, the procedure consisted of an induction phase, followed by the dependent measures, followed by check questions. In the induction phase, participants learned the event structure summarized in Figure 11:

> In a distant country, there is a group of people called the Favonians, with a unique culture involving practices such as *zorbing*, *kwerping*, *wanning*, *gorning*, and *lepping.*
>
> Sometimes some of these practices can trigger other practices, causing them to occur.
>
> Some of the Favonians' practices are steps involved in other practices. For example, in the United States, we have a practice of eating meals. Two of the steps involved in eating a meal are putting out the dishes and having dessert.
>
> In addition, some of the Favonians' practices are different sorts of other practices. For example, two different sorts of meals are breakfast and dinner.
>
> Favonians sometimes Favonians sometimes [*c*]. [*c*]ing is a step involved in [*A*]ing, and is a sort of [*B*]ing.
>
> [*a*]ing is another step involved in [*A*]ing.
>
> [*b*]ing is another sort of [*B*]ing.

That is, for the target event *c*, one event (*a*) was described as a partonomic match because it was another part of the same partonomic superordinate, whereas the other event (*b*) was described as a taxonomic match because it was another kind of the same taxonomic superordinate. The event categories *A*, *a*, *B*, *b*, and *c* were replaced with novel event category names (e.g., *kwerp*) as in Experiment 4, using two pseudorandom assignments. The order of describing the partonomic (*a*) and taxonomic (*b*) relations was counterbalanced. On the same screen as this information, participants completed a series of 16 true/false check questions to ensure that they encoded the event structure (e.g., "Kwerping is a sort of gorning").

On the next page, participants read that "Lee is a Favonian who [*c*]ed, [*a*]ed, and [*b*]ed." In Experiment 4A, participants then completed a *causality* question: "Lee's [*c*]ing was caused either by his [*a*]ing or by his [*b*]ing, but we aren't sure which. Which do you think is more likely to have caused Lee to [*c*]?" Answers were entered on a sliding scale from 1 ("Definitely [*a*]ing") to 9 ("Definitely [*b*]ing"). In Experiment 4B, participants completed a *similarity* question: "Think of how similar Lee's [*a*]ing is to Lee's [*c*]ing, and of how similar Lee's [*b*]ing is to Lee's [*c*]ing. Which do you think is more similar to Lee's [*c*]ing?" Answers were entered on a sliding scale from 1 ("[*a*]ing") to 9 ("[*b*]ing"). The order of listing *a* and *b* was counterbalanced in all of the above information and measures.

Results and Discussion

Responses were converted to a scale from -4 to 4, where negative scores corresponded to endorsements of the partonomic match (*a*) and positive scores corresponded to endorsements of the taxonomic match (*b*). Participants in Experiment 4A indicated that the partonomic match (*a*) was a more likely cause of the target event compared to the taxonomic match (*b*), leading to significantly negative scores (M = -0.84, SD = 2.66), $t(66) = -2.58$, $p = .012$, $d = -0.31$. However,

contrary to a similarity-based account, the *taxonomic* matches were considered more similar to the effect event in Experiment 4B, leading to significantly positive scores ($M = 0.66$, $SD = 2.77$), $t(71) = 2.03$, $p = .047$, $d = 0.24$. This led to a significant difference between the responses to the causality and similarity questions, $t(137) = 3.25$, $p = .001$, $d = 0.55$.[4]

This result shows that similarity-matching not only cannot explain the use of event structure cues in causal reasoning, but that event structure cues can even override similarity. This is particularly striking in light of the robust effects of similarity on causal judgments (Einhorn & Hogarth, 1986; LeBoeuf & Norton, 2012; Shultz & Ravinsky, 1977; Shweder, 1977; White, 2009). Of course, we are not suggesting that similarity-matching does not occur in causal judgment—it is culturally ubiquitous (Shweder, 1977), possibly innate (Michotte, 1963/1946), and often an adaptive constraint for constraining causal inference (Einhorn & Hogarth, 1986; Herschel, 2009/1830). However, when similarity and event structure cues are in conflict, event structure cues may often dominate.

General Discussion

Constraining the space of possible causes is a critical yet computationally difficult problem faced by people in a wide array of everyday tasks. We face this problem when we puzzle out the workings of some physical causal system, when we assign blame to others for things that have gone awry, when we plan interventions to alter the world, and when we predict

[4] Given the relatively high exclusion rate for this experiment, it is useful to note that the pattern of results is similar when all participants are included in the analysis. Although the causality effect in Experiment 4A does not reach significance ($M = -0.34$, $SD = 2.57$), $t(119) = -1.47$, $p = .145$, $d = -0.13$, the similarity effect in Experiment 4B remains significant ($M = 0.52$, $SD = 2.56$), $t(119) = 2.22$, $p = .029$, $d = 0.20$. Consequently, the difference between the causality and similarity questions remains significant, $t(238) = 2.60$, $p = .010$, $d = 0.34$, with the means on opposite sides of the scale midpoint.

the future. Yet, people are often able to solve this problem—*Peirce's problem* of hypothesis generation—in a seemingly effortless way.

Here, we showed that people use cues from the part-whole structure of events to constrain Peirce's problem. Participants were more likely to match high-level effects to high-level causes and low-level effects to low-level causes, demonstrating the *level-matching* effect (Experiments 1–4). However, they did not match low-level effects to low-level causes across a high-level event boundary, demonstrating the *boundary-blocking* effect (Experiments 1–3), and did not match high-level effects to high-level causes when the high-level events were not united in a partonomic structure (Experiment 3). These effects occurred in both diagnostic (Experiments 1, 3, and 4) and predictive reasoning (Experiment 2), and regardless of whether information about temporal order was present or absent, and whether a human agent had intervened at each step in the hierarchy (Experiments 1 and 2). Further, these effects were robust across both physical (Experiments 1 and 2) and social causal systems (Experiments 3 and 4), and were strong enough to trump resemblance-based causal reasoning (Experiment 4). In what follows, we consider explanations for these phenomena, and connect these findings to broader theoretical questions about causal reasoning and representation in cognitive, social, and developmental psychology.

Accounting for the Effects of Event Structure

In the introduction, we suggested why event structure cues are an adaptive strategy for identifying relatively plausible causal candidates. Events at the same hierarchical level are likely to stand in a specific (i.e., close to one-to-one) relationship with each other (Campbell, 2008; Woodward, 2010), and events from the same event cluster are likely to co-occur more robustly to changes in background conditions (Lewis, 1973; Woodward, 2006). Because specific and robust

causal relationships are more reliable predictors of future events and better control variables for bringing about desired effects, cues that help to realize these features would be useful heuristics to apply in causal inference.

Although we think that this account of event structure cues is simple and intuitive, several alternative possibilities remain that could explain at least a subset of our results. Here, we consider *construal level*, *similarity*, and *testimony* accounts of our findings.

Construal level. Events at relatively abstract levels of analysis (such as the high-level events in our experiments) are construed in more psychologically distant ways compared to events are more concrete levels of analysis (Trope & Liberman, 2003, 2010). In particular, high-level events may be seen as more *temporally* distant, which could prompt people to seek more temporally distant causes for them. This could lead to a level-matching effect.

However, construal level theory (CLT) cannot explain several other results in our studies. Most critically, Experiment 3 directly tested a CLT account by comparing high-high level-matching in event structures that had or lacked a unifying 'super-superordinate' event (see Figures 9A and 9B). Although the presence of this unifying event would not alter the relative construal level of the high-level events (because they are at the same level as each other), we found in Experiment 3 that it was essential for level-matching to occur. This prediction is consistent with our event structure framework, but cannot be explained by a CLT account.

Similarly, it is unclear how CLT would account for boundary-blocking, wherein low-level events are matched only when united by a common high-level event. Boundary-blocking differs from the effect of the 'super-superordinate' event from Experiment 3 in that the events in boundary-blocking all belong to a common hierarchy (in contrast to the structure in Figure 9B that has no unifying structure between the two branches). However, CLT would seem to imply

that all low-level events would be construed at the same (concrete and psychologically proximal) level, and would not predict any difference in matching within or across a superordinate event, unless supplemented with additional principles.

Finally, a CLT account appears to be in conflict with the results of Experiment 4, wherein people preferred to match an effect to a partonomically matched rather than a hierarchically matched cause. Both taxonomic and partonomic structures vary in abstractness (indeed, manipulations in the CLT literature often involve taxonomic structures; Trope & Liberman, 2010). Thus, CLT seems to give no reason to prefer either a partonomic or a taxonomic match when given the choice. Yet, our participants preferred the partonomic match, consistent with the idea that the critical factor for their inferences was event structure rather than level of abstraction. Of course, construal level plays an important role in guiding inferences in other contexts, such as decision-making and person perception (e.g., Trope & Liberman, 2003, 2010), and has been shown to modulate other causal reasoning processes, such as the relative focus on the causes or effects of events (Rim, Hansen, & Trope, 2012). In the cases studied here, however, event structure was a much stronger factor in narrowing down potential causes and effects than was construal level.

Similarity-matching. One natural account of the level-matching phenomenon is that participants are simply matching similar effects to similar causes, and vice versa, as they have been shown to do in many laboratory experiments (Einhorn & Hogarth, 1986; White, 2009) and field studies (Frazer, 1959; Shweder, 1977). Boundary-blocking could potentially be accounted for as well, on the assumption that events unified by a superordinate would share more features in common than those that cross superordinates.

A similarity account, however, does not supplant our event structure account for two reasons. First, the similarity account is too unconstrained without additional assumptions. Similarity can be computed along many dimensions (Goodman, 1955; Medin, Goldstone, & Gentner, 1993), and principles similar to our event structure cues are necessary to constrain similarity in a way that would produce our results. Second, despite this lack of constraint, similarity is empirically inadequate for explaining these results. In Experiment 4, we found that taxonomic level-matches were seen as more similar than partonomic level-matches, yet people still preferred the partonomic match in making causal judgments.

Testimony-based inference. The current experiments used 'described' rather than 'experienced' events to test the use of event structure cues. That is, we tested participants' inferences about event structures wherein the events were pre-individuated for the participant. This method has several advantages over a less constrained method wherein participants must themselves individuate the causal structure—it allowed us greater control over how participants are interpreting the stimuli, minimized noise due to individual idiosyncrasies in event parsing, and gave us precision for testing our specific hypotheses. Moreover, in other causal reasoning tasks, inferences from described and experienced situations tend to align (e.g., Griffiths & Tenenbaum, 2005; Wasserman, 1990).

Despite these advantages, the use of described events raises the question of what set of assumptions participants were using for making causal inferences from these descriptions. We have argued that participants are making these inferences based on hierarchical cues in the event structure itself, which would be similar to inferences from experienced events. However, perhaps participants are using assumptions about how the descriptions were generated by an individual who perceived the action sequence. People rely on a number of cues for individuating events,

which include spatial, temporal, and object cues as well as causal cues (Buchsbaum et al., 2009; Zacks et al., 2009). Thus, if causal cues were used to individuate the structures as described by informed speakers, participants could instead be relying on assumptions about speakers' causal knowledge, rather than assumptions about the event structure itself. Put differently, participants might see the event descriptions as coming from informed speakers who used causal knowledge to individuate the event structure, and participants might then see their task as attempting to recover or decode the original causal structure that led to that pattern of described events.

Although testimony-based inferences are an interesting possibility in many contexts, they are unlikely to fully account for our current results for both conceptual and empirical reasons. First, we used a variety of causal systems (physical systems such as chemical reactions and machine operations, more abstract systems such as computer subroutines, and social systems such as the practices of the Favonians), and the plausibility of this explanation varies considerably across stimuli. For some sorts of physical systems, such as the chemical reaction vignette, participants might assume that the speaker has expert knowledge and would have used primarily causal information for individuating events. In other cases, however, it seems more likely that other structure cues would be primarily used for individuation. In our machine vignettes, for example, the events were the operation of different *parts* of the machine, suggesting that the primary cue for individuation would be object-based rather than causal. And in our computer vignettes, the events were subroutines of computer programs which have object boundaries defined by the user. Despite the varying plausibility of the testimony explanation across these cover stories, we nonetheless found robust results across all of our experiments.

A related conceptual problem with a testimony-based account is that it must make further assumptions about what participants assume about speakers' causal knowledge, and about how

that knowledge is translated into event structures. That is, on a testimony-based account, participants must assume that the descriptions were generated by speakers who have access to the causal structure of the situations. But how did those speakers acquire access to the causal structure? In some cases, speakers could plausibly be assumed to be experts with access to specialized knowledge (e.g., in the chemistry story) that could be used for individuation. But in other cases (e.g., the social practices of the Favonians) domain experts with this sort of specialized knowledge are much less plausible. In cases like the Favonians, the speaker would need some independent way of learning the causal structure *other than the event structure* in order for a testimony-based account to avoid circularity. Further, participants would also need a set of assumptions guiding how the speaker used their (assumed) causal knowledge to individuate the event structure. These assumptions would have to include at least some notion that causal links trigger event segmentation at multiple levels of granularity (to justify the level-matching inference) and further segmentation principles to justify the boundary-blocking inference. Although it is possible that participants make these detailed assumptions about the speakers' causal knowledge and event individuation capacities, the need to impute these rather complex beliefs to participants counts against the plausibility of the testimony-based account.

Finally, there is empirical reason to doubt the testimony explanation. Causal cues are used for individuating both high-level and low-level events, and are used somewhat more robustly for individuating high-level events (Zacks et al., 2009). If people level-match because they assume that the descriptions were generated by speakers using causality to individuate events, one would therefore expect the level-matching effect to be of similar strength for low-level and high-level events, or perhaps somewhat stronger for high-level events. This was not what we observed in the current studies. Across all experiments where both high-high and low-

low level-matching were measured, the low-low level-matching effect was stronger (see Figures 2, 5, and 10). This result fits nicely into our framework, as people are likely to assume that low-level events are seen as more precise control variables than high-level events (i.e., easier to intervene on precisely to bring about a desired effect). However, it would be difficult to explain on the testimony account.

Thus, although the current experiments do not rule out a testimony-based explanation in all possible cases, such an explanation is inconsistent with our observed results and requires further assumptions to make plausible.

Toward an Account of Causal-Temporal Inference

Temporal cues may be among the most important source of information we have for narrowing down the hypothesis space of causal explanations. In particular, people use the principles of temporal *priority* (causes always precede their effects in time) and temporal *contiguity* (causes and effects tend to be adjacent in time) to constrain the space of possible causes. Both adults and children use these temporal principles for causal inference (Lagnado & Sloman, 2006; Mendelson & Shultz, 1976; Rottman & Keil, 2012; Rottman, Kominsky, & Keil, 2013), and the visual system is finely attuned to the temporal parameters of displays in perceiving causality (Michotte, 1963/1946).

However, temporal priority and contiguity alone underdetermine which event to entertain as a causal candidate. These cues may point toward ‘the event’ that preceded the effect, but these cues do not demarcate what portion of time constitutes the preceding event. Principles of event individuation are necessary to more fully constrain Peirce’s problem. Our event hierarchy framework provides a set of principles which, in conjunction with event parsing principles (e.g.,

Buchsbaum et al., 2009; Zacks et al., 2009) and temporal priority and contiguity (Lagnado & Sloman, 2006; Mendelson & Shultz, 1976), can provide such additional constraints.

Event hierarchy influences can help to explain some recent findings concerning the relationship between temporal delay and causal inferences. Although there is a general tendency for the perception of a causal relationship to weaken over increasingly long intervals between two events (e.g., Michotte, 1963/1946), other research has shown that this relationship is not always monotonic—in some situations, longer delays actually *facilitate* causal inferences, when such delays are in accord with prior expectations (Buehner & May, 2002; Buehner & McGregor, 2006). In other words, people do not have invariant expectations about the delay between cause and effect in terms of absolute time, but have different expectations depending on the contents of those events. Expectations about temporal delays may not be conceptualized in terms of *time* as such, but rather in terms of *event units*, which vary according to the hierarchical position of the cause and effect. Since high-level events by definition take longer than their lower-level subordinates, and they are expected to have higher-level causes, one would also expect a longer delay between the initiation of cause and initiation of effect for higher-level causes.

We conducted an experiment as a first test of this idea. When historical or natural events (e.g., the Hundred Years' War, or the formation of a galaxy) were described at a relatively high level of an event hierarchy (i.e., as containing many sub-parts), participants tended to think that the causes of those events are perturbed back further in time, compared to when those same events were described at a relatively low level of an event hierarchy (i.e., as contained within a larger event). That is, the expected temporal delay between cause and effect is modulated by the level of the event hierarchy at which the effect is described. This finding should be taken as preliminary, since it could also be explained by other mechanisms (e.g., higher-level construal

resulting in people thinking of the event as more temporally distant; Trope & Liberman, 2003). Nonetheless, this result shows that hierarchical structure influences delay expectations, and is at least suggestive of the possibility that these expectations are conceptualized in terms of event units.

Origins of Causal Knowledge

By adulthood, we have considerable expertise at inferring the causes of particular events because we have extensive causal knowledge. But independent of the extent to which we have core knowledge of basic causal principles (e.g., Carey, 2009; Leslie, 1994; Setoh, Wu, Baillargeon, & Gelman, 2013; Spelke, 1990), we surely are not born with *specific* causal knowledge about particular social conventions (e.g., waiting in line), artifacts (e.g., telephones), or natural kinds (e.g., duck-billed platypi). Statistical learning alone—that is, the use of co-occurrence among events to rationally infer causal relationships—is not sufficient to get causal learning off the ground because constraints are needed to specify what events to sample. For example, Sobel and Kirkham (2006) investigated inferences about the causal properties of a "blicket detector," demonstrating that 24-month-old children (using a forced-choice measure) and even 8-month-old infants (using an eyetracking measure) have rich abilities to marshal co-occurrence data to perform sophisticated statistical inferences. However, the children could not have made those inferences without antecedent understanding of how to parse their experience with the blicket detector into discrete events (e.g., this block is placed on the detector, this light goes on, this second block is placed on the detector, etc.).

Given that even 10-month-old infants can parse experience into discrete events (Baldwin, Baird, Saylor, & Clark, 2001), we conjecture that this ability is likely to underlie early-emerging causal competences (e.g., Gweon & Schulz, 2011; Leslie & Keeble, 1987; Sobel & Kirkham,

2006). Young children, like adults, may therefore use event structure to constrain causal inference, and perhaps rely on structure to an even greater extent because of their scarcer mechanistic and statistical knowledge. We know of no infant studies that have directly tested the relationship between event representations and causal understanding, but related abilities are present in 3-year-old children, who can make sophisticated inferences from event structure to choose which causal action sequences to imitate (Buchsbaum, Gopnik, Griffiths, & Shafto, 2011). We think it is plausible that young children and perhaps infants use principles similar to those demonstrated in the current studies to constrain Peirce's problem at a time when background knowledge is especially limited, and widely applicable cues like event structure are likely to be most critical.

Conclusion

Despite its massive multilayered complexity, we usually navigate the physical and social world with ease and often infer causal regularities in a seemingly effortless manner. A variety of competences seem to underlie this ability, including statistical learning, rapid application of prior knowledge, and automatic perceptual processing. Because the sense data of any particular experience underdetermines the causal structure of a particular episode, we must often rely on cues such as temporal priority and contiguity to make causal inferences. Here, we have shown that event structure—the organization of experience by the cognitive system into discrete events that are hierarchically organized—can be used as a cue to constrain causal inference, with events at a given level of a hierarchy causally matched to other events at that level (the *level-matching* effect) and specifically to events that are conceptualized as parts of the same superordinate event (the *boundary-blocking* effect). Because event structure is processed automatically and based in part on bottom-up perceptual cues, event structure information is available 'for free' or a very

low cost even in novel situations. Event structure cues therefore have the potential to powerfully constrain causal inferences in unfamiliar environments and under speeded real time conditions, and to explain our ability to track useful causal patterns in the world.

Acknowledgements

We thank Andy Jin and Greeshma Rajeev-Kumar for assistance in developing stimuli, Joshua Knobe and Laurie Santos for comments on an earlier version of this manuscript, the members of the Cognition and Development lab for their suggestions, and an audience at Yale University for helpful discussion. This research was funded by grant R37-HD23922 from the National Institutes of Health, awarded to the second author.

References

Abbott, V., Black, J. B., & Smith, E. E. (1985). The representation of scripts in memory. *Journal of Memory and Language*, *24*, 179–199.

Ahn, W., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition*, *54*, 299–352.

Alicke, M. D. (1992). Culpable causation. *Journal of Personality and Social Psychology*, *63*, 368–378.

Apperly, I. A. (2010). *Mindreaders: The cognitive basis of "theory of mind."* New York, NY: Psychology Press.

Baldwin, D. A., Baird, J. A., Saylor, M. M., & Clark, M. A. (2001). Infants parse dynamic action. *Child Development*, *72*, 708–717.

Baldwin, D., Andersson, A., Saffran, J., & Meyer, M. (2008). Segmenting dynamic human action via statistical structure. *Cognition*, *106*, 1382–1407.

Buchsbaum, D., Griffiths, T. L., Gopnik, A., & Baldwin, D. (2009). Learning from actions and their consequences: Inferring causal variables from continuous sequences of human action. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2493–2498). Austin, TX: Cognitive Science Society.

Buchsbuam, D., Canini, K. R., & Griffiths, T. L. (2011). Segmenting and recognizing human action using low-level video features. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society* (pp. 3162–3167). Austin, TX: Cognitive Science Society.

Buchsbaum, D., Gopnik, A., Griffiths, T. L., & Shafto, P. (2011). Children's imitation of causal action sequences is influenced by statistical and pedagogical evidence. *Cognition*, *120*, 331–340.

Buehner, M. J., & May, J. (2002). Knowledge mediates the timeframe of covariation assessment in human causal induction. *Thinking & Reasoning*, *8*, 269–293.

Buehner, M. J., & McGregor, S. (2006). Temporal delays can facilitate causal attribution: Towards a general timeframe bias in causal induction. *Thinking & Reasoning*, *12*, 353–378.

Campbell, J. (2008). Interventionism, control variables and causation in the qualitative world. *Philosophical Issues*, *18*, 426–445.

Carey, S. (2009). *The origin of concepts*. Oxford, UK: Oxford University Press.

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

Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. *Cognitive Psychology*, *17*, 391–416.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*, 353–380.

Dunning, D., Griffin, D. W., Milojkovic, J. D., & Ross, L. (1990). The overconfidence effect in social prediction. *Journal of Personality and Social Psychology*, *58*, 568–581.

Einhorn, H. J., & Hogarth, R. M. (1986). Judging probable cause. *Psychological Bulletin*, *99*, 3–19.

Frazer, J. G. (1959). *The new golden bough*. New York, NY: Criterion Books.

Fugelsang, J. A., Stein, C. B., Green, A. E., & Dunbar, K. N. (2004). Theory and data interactions of the scientific mind: Evidence from the molecular and the cognitive laboratory. *Canadian Journal of Experimental Psychology*, *58*, 86–95.

Goldvarg, E., & Johnson-Laird, P. N. (2001). Naive causality: A mental model theory of causal meaning and reasoning. *Cognitive Science*, *25*, 565– 610.

Goodman, N. (1955). *Fact, fiction, and forecast*. Cambridge, MA: Harvard University Press.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*, 334–384.

Gweon, H., & Schulz, L. (2011). 16-month-olds rationally infer causes of failed actions. *Science*, *332*, 1524.

Hart, H. L. A., & Honoré, T. (1985). *Causation in the law* (2nd ed.). Oxford, UK: Clarendon.

Hastie, R. (1984). Causes and effects of causal attribution. *Journal of Peresonality and Social Psychology*, *46*, 44–56.

Heit, E. (2000). Properties of inductive reasoning. *Psychonomic Bulletin & Review*, *7*, 569–592.

Herschel, J. F. W. (2009). *A preliminary discourse on the study of natural philosophy*. Cambridge, UK: Cambridge University Press.

Hill, A. B. (1965). The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine, 58*, 295–300.

Hilton, D. J., McClure, J., & Sutton, R. M. (2010). Selecting explanations from causal chains: Do statistical principles explain preferences for voluntary causes? *European Journal of Social Psychology*, *40*, 383–400.

Hoover, K. D. (2001). *Causality in macroeconomics*. Cambridge, UK: Cambridge University Press.

Johnson, S. G. B., & Ahn, W. (*in press*). Causal networks or causal islands? The representation of mechanisms and the transitivity of causal judgment. *Cognitive Science*.

Johnson, S. G. B., Jin, A., & Keil, F. C. (2014). Simplicity and goodness-of-fit in explanation: The case of intuitive curve-fitting. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 701–706). Austin, TX: Cognitive Science Society.

Johnson, S. G. B., Johnston, A. M., Toig, A. E., & Keil, F. C. (2014). Explanatory scope informs causal strength inferences. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 2453–2458). Austin, TX: Cognitive Science Society.

Johnson, S. G. B., & Keil, F. C. (2014). Pluralism in causal learning strategies. *Manuscript under review*.

Johnson, S. G. B., Rajeev-Kumar, G., & Keil, F. C. (2014). Inferred evidence in latent scope explanations. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 707–712). Austin, TX: Cognitive Science Society.

Kelley, H. H. (1973). The processes of causal attribution. *American Psychologist*, *28*, 107–128.

Kendler, K. S. (2005). "A gene for...": The nature of gene action in psychiatric disorders. *American Journal of Psychiatry*, *162*, 1243–1252.

Khemlani, S. S., Sussman, A. B., & Oppenheimer, D. M. (2011). Harry Potter and the sorcerer's scope: Latent scope biases in explanatory reasoning. *Memory & Cognition*, *39*, 527–535.

Kunda, Z., & Thagard, P. (1996). Forming impressions from stereotypes, traits, and behaviors: A parallel-constraint-satisfaction theory. *Psychological Review*, *103*, 284–308.

Lagnado, D. A., & Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition*, *108*, 754–770.

Lagnado, D. A., & Sloman, S. A. (2006). Time as a guide to cause. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 451–460.

LeBoeuf, R. A., & Norton, M. I. (2012). Consequence-cause matching: Looking to the consequences of events to infer their causes. *Journal of Consumer Research*, *39*, 128–141.

Leslie, A. M. (1994). ToMM, ToBy, and Agency: Core architecture and domain specificity. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 119–148). Cambridge, UK: Cambridge University Press.

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

Lewis, D. (1973). Causation. *The Journal of Philosophy*, *70*, 556–567.

Lewis, D. (2000). Causation as influence. *The Journal of Philosophy*, *97*, 182–197.

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

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, *55*, 232–257.

Lombrozo, T. (2010). Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, *61*, 303–32.

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

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, *100*, 254–278.

Mendelson, R., & Shultz, T. R. (1976). Covariation and temporal contiguity as principles of causal influence in young children. *Journal of Experimental Child Psychology*, *22*, 408–412.

Michotte, A. (1963). *The perception of causality*. (T. R. Miles & E. Miles, Trans.). New York: Basic Books. (Original work published 1946.)

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice Hall.

Newtson, D. (1973). Attribution and the unit of perception of ongoing behavior. *Journal of Personality and Social Psychology*, *28*, 28–38.

Newtson, D., Hairfield, J., Bloomingdale, J., & Cutino, S. (1987). The structure of action and interaction. *Social Cognition*, *5*, 191–237.

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

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

Peirce, C. S. (1997). *Pragmatism as a principle and method of right thinking: The 1903 Harvard lectures on pragmatism*. (P. A. Turrisi, Ed.). Albany, NY: State University of New York Press. (Original work published 1903.)

Read, S. J., & Marcus-Newhall, A. (1993). Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology*, *65*, 429–447.

Rim, S., Hansen, J., & Trope, Y. (2013). What happens why? Psychological distance and focusing on causes versus consequences of events. *Journal of Personality and Social Psychology*, *104*, 457–472.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 382–439.

Rottman, B. M., & Keil, F. C. (2012). Causal structure learning over time: Observations and interventions. *Cognitive Psychology*, *64*, 93–125.

Rottman, B. M., Kominsky, J. F., & Keil, F. C. (2014). Children use temporal cues to learn causal directionality. *Cognitive Science*, *38*, 489–513.

Rozin, P., Millman, L., & Nemeroff, C. (1986). Operation of the laws of sympathetic magic in disgust and other domains. *Journal of Personality and Social Psychology*, *50*, 703–712.

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton, NJ: Princeton University Press.

Setoh, P., Wu, D., Baillargeon, R., & Gelman, R. (2013). Young infants have biological expectations about animals. *Proceedings of the National Academy of Sciences*, *110*, 15937–15942.

Shultz, T. R., & Ravinsky, F. B. (1977). Similarity as a principle of causal inference. *Child Development*, *48*, 1552–1558.

Shweder, R. A. (1977). Likeness and likelihood in everyday thought: Magical thinking in judgments about personality. *Current Anthropology*, *18*, 637–658.

Sloman, S. A., Barbey, A. K., & Hotaling, J. M. (2009). A causal model theory of the meaning of *Cause*, *Enable*, and *Prevent*. *Cognitive Science*, *33*, 21–50.

Sloman, S. A., & Hagmayer, Y. (2006). The causal psycho-logic of choice. *Trends in Cognitive Sciences*, *10*, 407–412.

Sobel, D. M., & Kirkham, N. Z. (2006). Blickets and babies: The development of causal reasoning in toddlers and infants. *Developmental Psychology*, *42*, 1103–1115.

Spelke, E. S. (1990). Principles of object perception. *Cognitive Science*, *14*, 29–56.

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search*. New York, NY: Springer.

Strevens, M. (2008). *Depth: An account of scientific explanation*. Cambridge, MA: Harvard University Press.

Strickland, B., Silver, I., & Keil. F. C. (2014). The texture of causes and effects: Domain specificity in causal reasoning. *Manuscript under review*.

Trope, Y., & Liberman, N. (2003). Temporal construal. *Psychological Review*, *110*, 403–421.

Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review*, *117*, 440–463.

Tversky, B., & Hemenway, K. (1984). Objects, parts, and categories. *Journal of Experimental Psychology: General*, *113*, 169–193.

Uleman, J. S., Saribay, S. A., & Gonzalez, C. M. (2008). Spontaneous inferences, implicit impressions, and implicit theories. *Annual Review of Psychology*, *59*, 329–360.

Waldmann, M. R., & Holyoak, K. J. (1992). Predictive and diagnostic learning within causal models: Asymmetries in cue competition. *Journal of Experimental Psychology: General*, *121*, 222–236.

Wasserman, E. A. (1990). Attribution of causality to common and distinctive elements of compound stimuli. *Psychological Science*, *1*, 298–302.

Wells, G. L. (1982). Attribution and reconstructive memory. *Journal of Experimental Social Psychology*, *18*, 447–463.

White, P. A. (2009). Property transmission: An explanatory account of the role of similarity information in causal inference. *Psychological Bulletin*, *135*, 774–793.

Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General*, *136*, 82–111.

Woodward, J. (2005). *Making things happen: A theory of causal explanation*. Oxford, UK: Oxford University Press.

Woodward, J. (2006). Sensitive and insensitive causation. *Philosophical Review*, *115*, 1–50.

Woodward, J. (2010). Causation in biology: Stability, specificity, and the choice of levels of explanation. *Biology and Philosophy*, *25*, 287–318.

Zacks, J. M. (2004). Using movement and intentions to understand simple events. *Cognitive Science*, *28*, 979–1008.

Zacks, J. M., Speer, N. K., & Reynolds, J. R. (2009). Segmentation in reading and film comprehension. *Journal of Experimental Psychology: General*, *138*, 307–327.

Zacks, J. M., & Tversky, B. (2001). Event structure in perception and conception. *Psychological Bulletin*, *127*, 3–21.

Zacks, J. M., Tversky, B., & Iyer, G. (2001). Perceiving, remembering, and communicating structure in events. *Journal of Experimental Psychology: General*, *130*, 29–58.

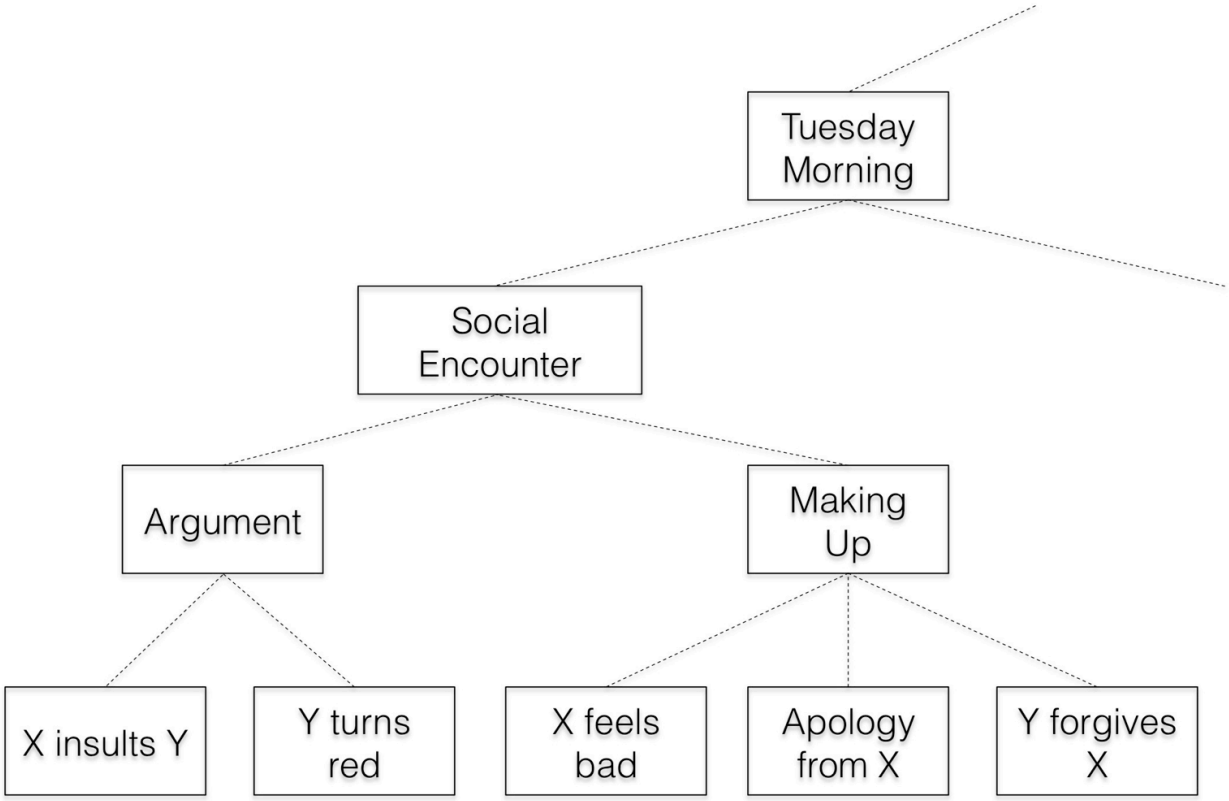

Figure 1. An example of an event hierarchy representing a social encounter between persons X and Y. The two events on the bottom left ("X insults Y" and "Y turns red") constitute a cluster because they are subordinates of the same higher-level event ("Argument"), as are the three events on the bottom right ("X feels bad," "Apology from X," and "Y forgives X") because they are subordinates of "Making Up." Additionally, "Argument" and "Making Up" form a higher-order cluster because they are both subordinates of "Social Encounter."

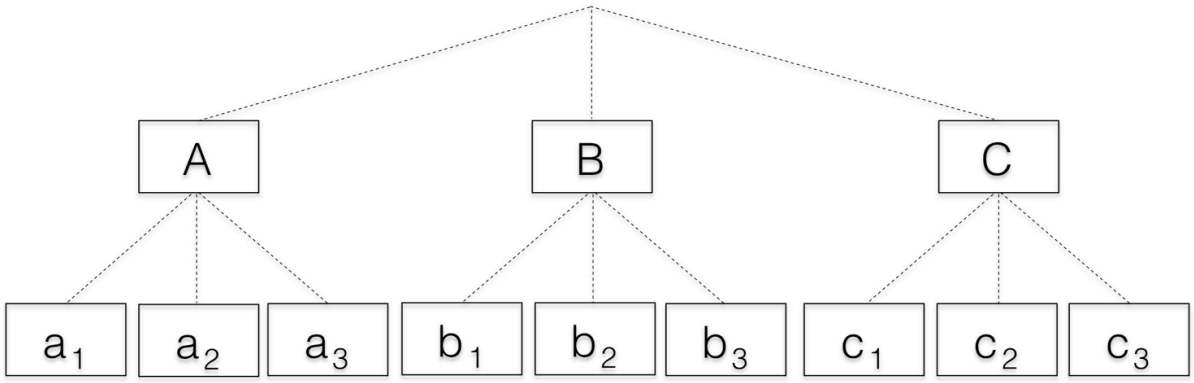

Figure 2. Event structure used in Experiments 1 and 2. Low-level events are represented by lowercase letters, and their higher-level superordinates are represented by uppercase letters.

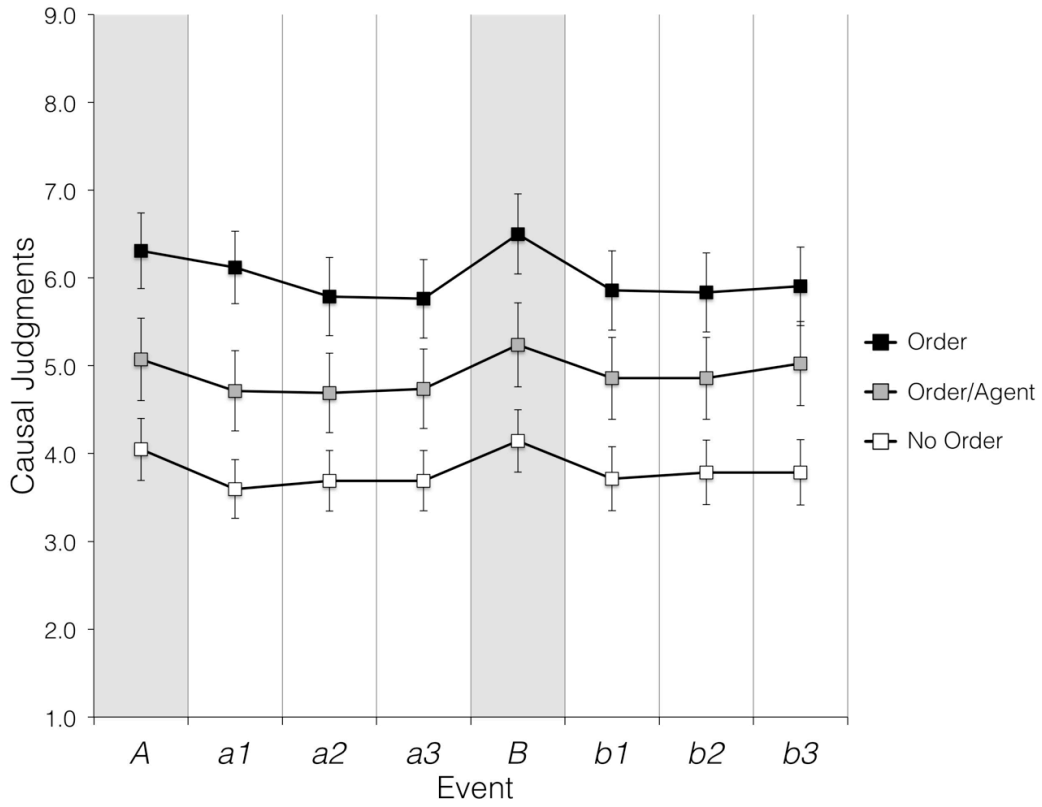

Figure 3. Ratings of causes of event *C* in Experiment 1 on a 9-point scale. Participants were not asked for judgments about events that were not logically distinct from *C* (i.e., *c1*, *c2*, and *c3*). Error bars represent one standard error of the mean.

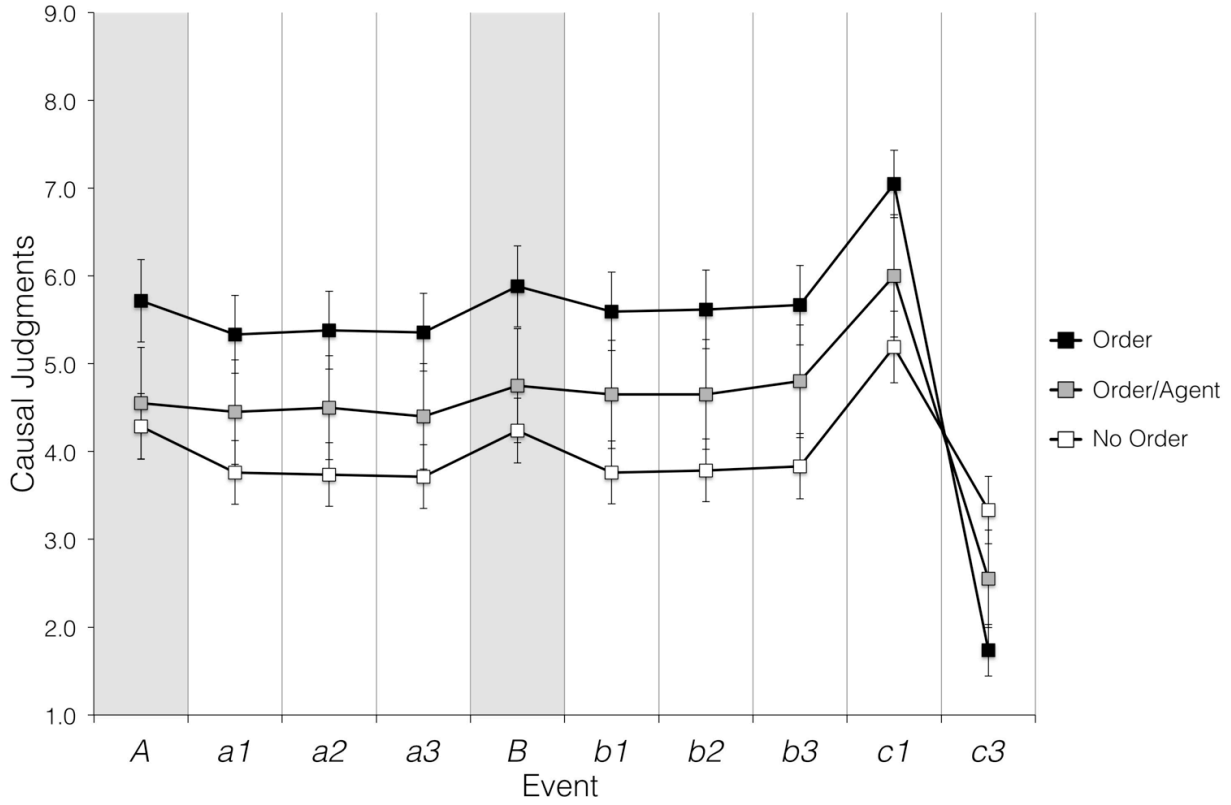

Figure 4. Ratings of causes of event *c2* in Experiment 1 on a 9-point scale. Participants were not asked for judgments about events that were not logically distinct from *c2* (i.e., *C*). Error bars represent one standard error of the mean.

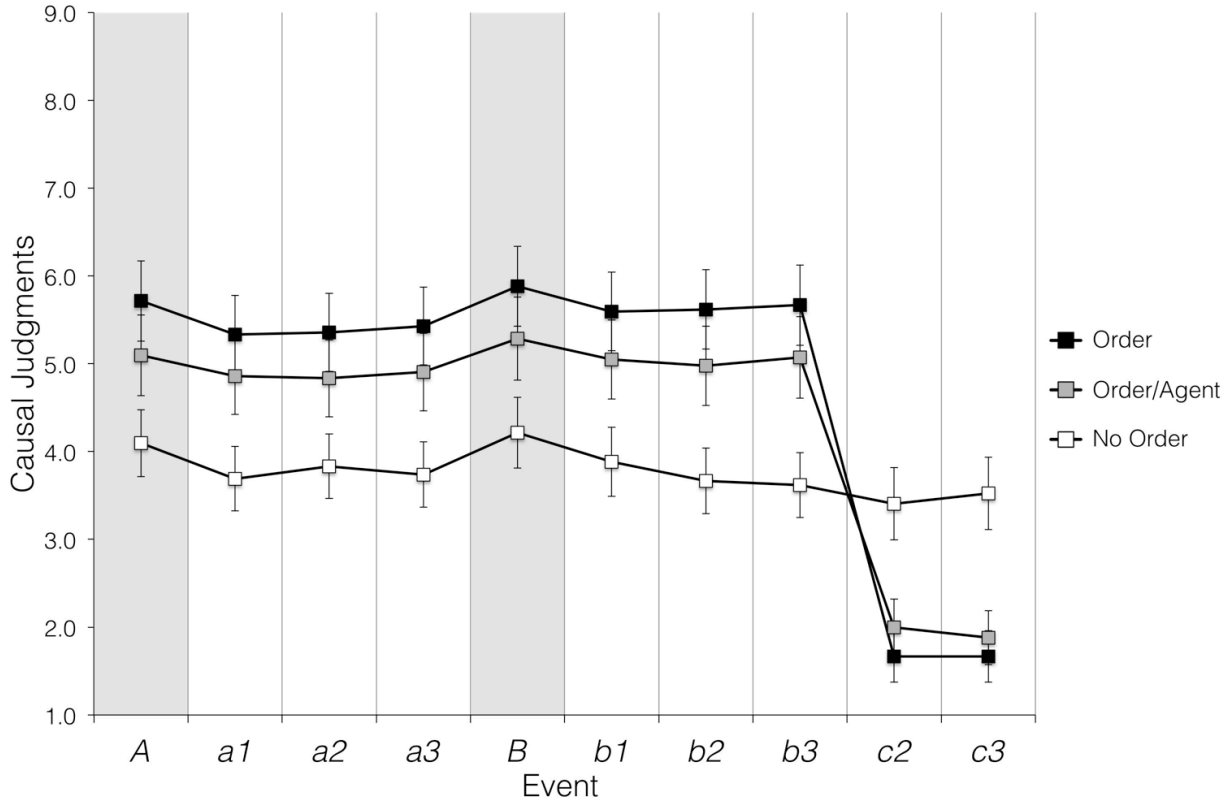

Figure 5. Ratings of causes of event *c1* in Experiment 1 on a 9-point scale. Participants were not asked for judgments about events that were not logically distinct from *c1* (i.e., *C*). Error bars represent one standard error of the mean.

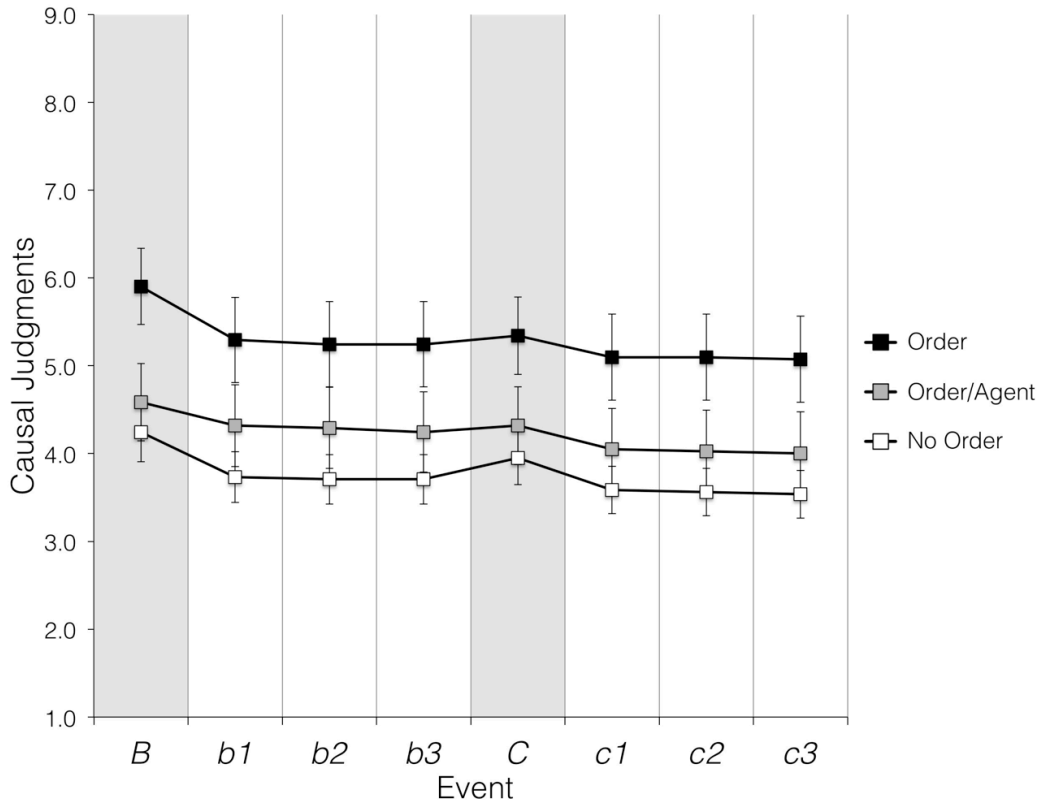

Figure 6. Ratings of effects of event *A* in Experiment 2 on a 9-point scale. Participants were not asked for judgments about events that were not logically distinct from *A* (i.e., *a1, a2*, and *a3*). Error bars represent one standard error of the mean.

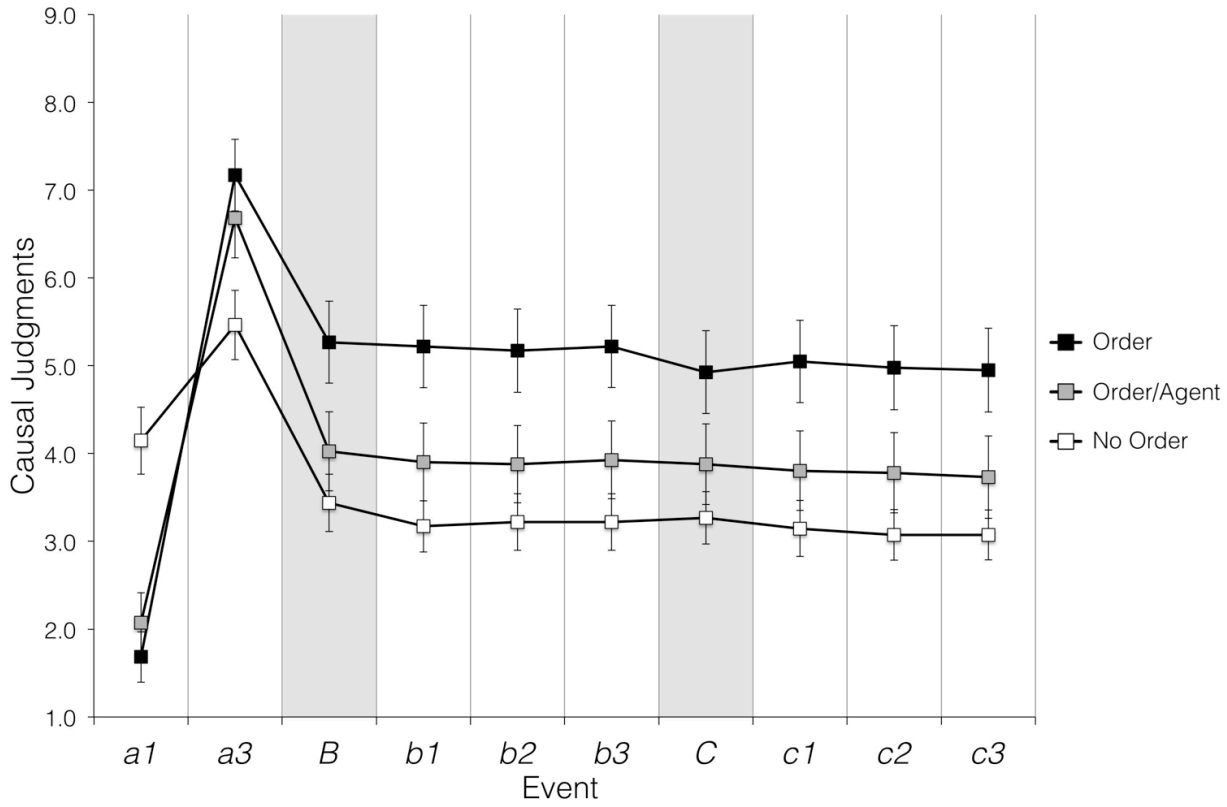

Figure 7. Ratings of effects of event *a2* in Experiment 2 on a 9-point scale. Participants were not asked for judgments about events that were not logically distinct from *a2* (i.e., *A*). Error bars represent one standard error of the mean.

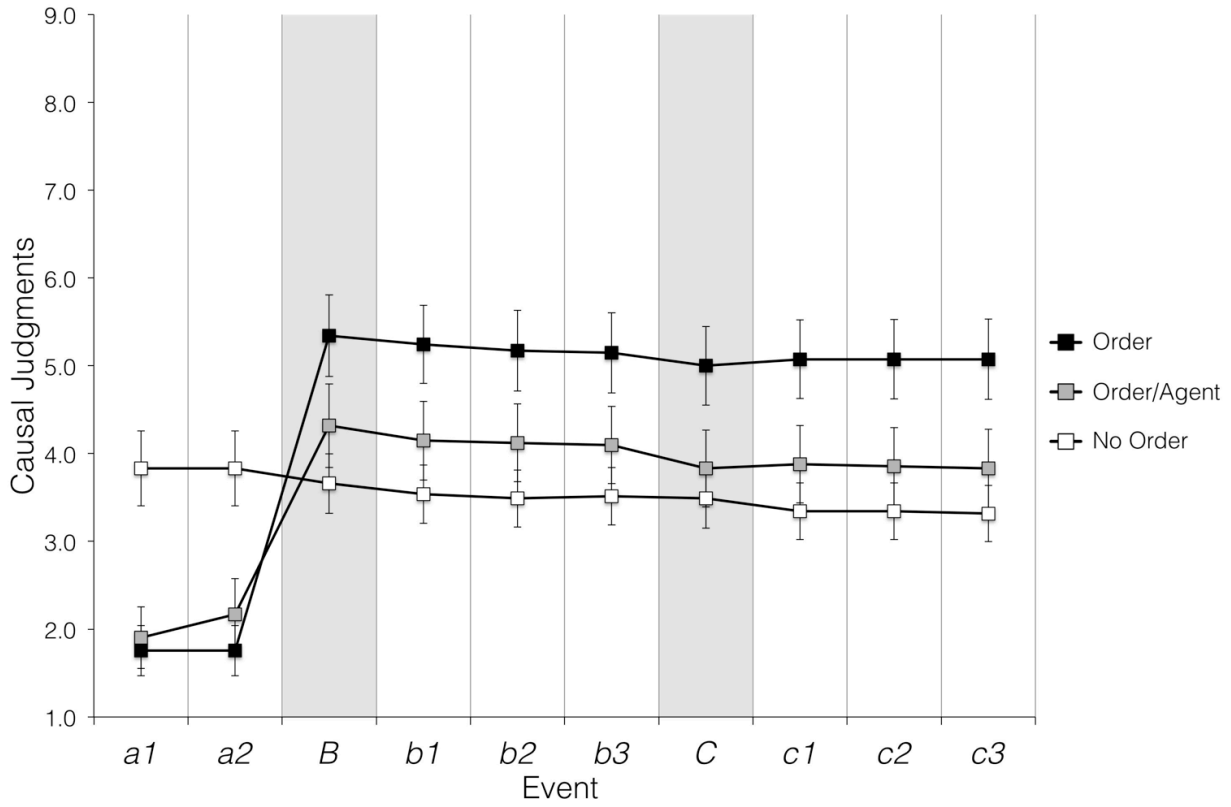

Figure 8. Ratings of effects of event *a3* in Experiment 2 on a 9-point scale. Participants were not asked for judgments about events that were not logically distinct from *a3* (i.e., *A*). Error bars represent one standard error of the mean.

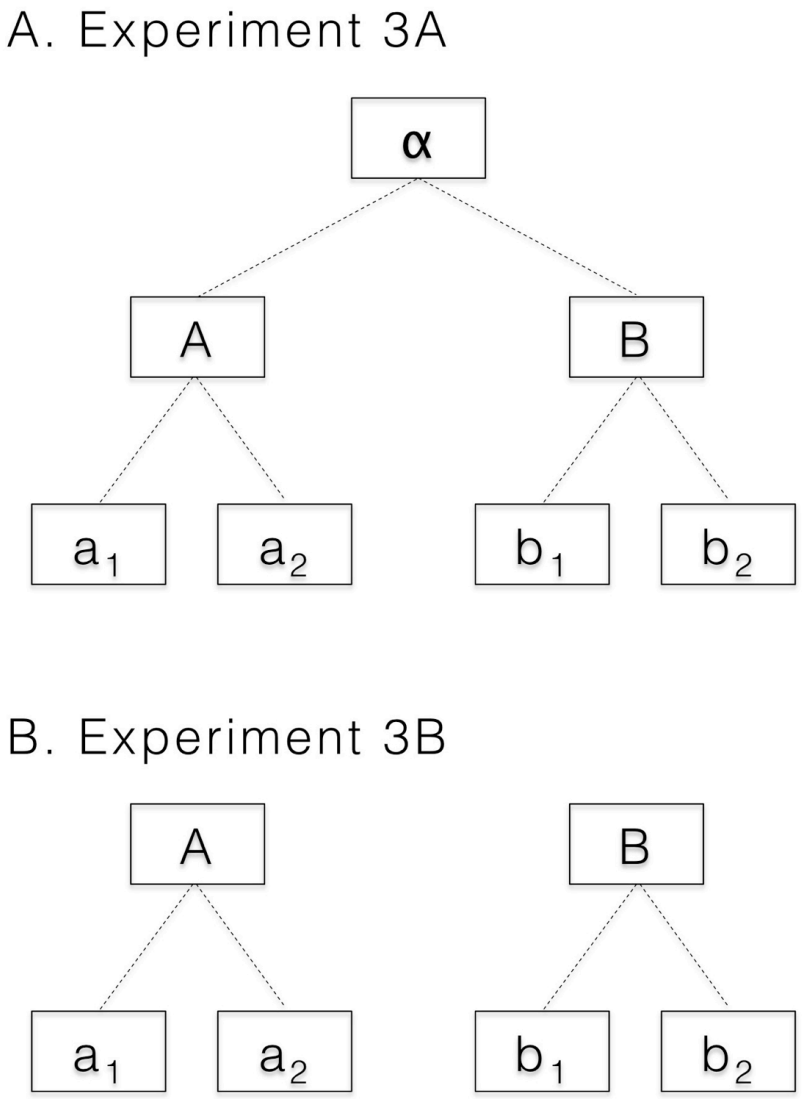

Figure 9. Event structure used in Experiments 3A (upper panel) and 3B (lower panel). Low-level events are represented by lowercase letters, and their higher-level superordinates are represented by uppercase letters. The 'super-superordinate' event in the upper panel is represented by α.

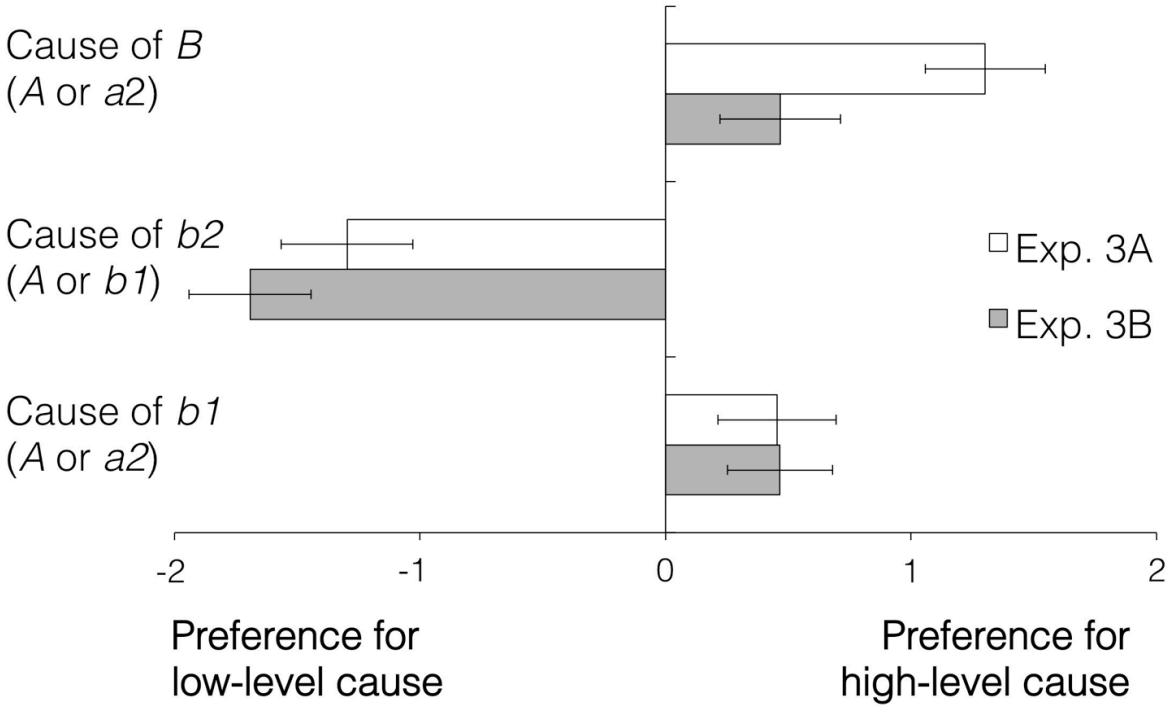

Figure 10. Results of Experiments 3A and 3B. Negative scores correspond to a preference for the low-level cause, and positive scores to a preference for the high-level cause. Error bars represent one standard error of the mean.

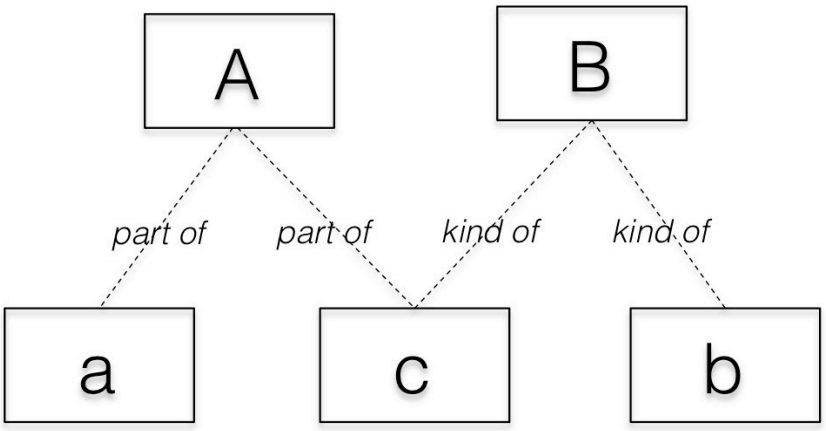

Figure 11. Event structure used in Experiment 4. Low-level events are represented by lowercase letters, and their higher-level superordinates are represented by uppercase letters. *A* is a partonomic superordinate for *a* and *c*, and *B* is a taxonomic superordinate for *b* and *c*.