COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science (2015) 1–36

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12213

Causal Networks or Causal Islands? The Representation of Mechanisms and the Transitivity of Causal Judgment

Samuel G. B. Johnson, Woo-kyoung Ahn

Department of Psychology, Yale University

Received 16 August 2013; received in revised form 1 June 2014; accepted 23 September 2014

Abstract

Knowledge of mechanisms is critical for causal reasoning. We contrasted two possible organizations of causal knowledge—an interconnected causal *network*, where events are causally connected without any boundaries delineating discrete mechanisms; or a set of disparate mechanisms —causal *islands*—such that events in different mechanisms are not thought to be related even when they belong to the same causal chain. To distinguish these possibilities, we tested whether people make *transitive* judgments about causal chains by inferring, given *A causes B* and *B causes C*, that *A causes C*. Specifically, causal chains schematized as one chunk or mechanism in semantic memory (e.g., exercising, becoming thirsty, drinking water) led to transitive causal judgments. On the other hand, chains schematized as multiple chunks (e.g., having sex, becoming pregnant, becoming nauseous) led to intransitive judgments despite strong intermediate links (Experiments 1–3). Normative accounts of causal intransitivity could not explain these intransitive judgments (Experiments 4 and 5).

Keywords: Causal mechanisms; Knowledge representation; Causal reasoning; Transitive inference

1. Introduction

Causal inference underlies our ability to predict the future, to explain the past, and to plan interventions on our environment. To make these inferences, humans rely on a remarkable ability to narrow the set of potential causes to a manageable size (Ahn & Kalish, 2000; Johnson & Keil, 2014; Lagnado, Waldmann, Hagmayer, & Sloman, 2007; Peirce, 1997/1903). One critical cue for narrowing this hypothesis space is knowledge of *causal mechanisms*—that is, knowledge of processes that reliably lead from causes to effects (e.g., Ahn, Kalish, Medin, & Gelman, 1995; Bullock, Gelman, & Baillargeon, 1982; Shultz, 1982). In this study, we examine how causal mechanisms are mentally

Correspondence should be sent to Sam Johnson, Department of Psychology, Yale University, 2 Hillhouse Ave., New Haven, CT 06520. E-mail: samuel.johnson@yale.edu

represented—as interconnected networks of influence, or as isolated causal schemas decontextualized from other mechanisms. We use the phenomenon of intransitive causal judgment (i.e., *A* causes *B*, *B* causes *C*, but *A* does not cause *C*) to argue that mechanisms are represented as relatively isolated chunks and to show that this representational format has downstream consequences for causal judgment.

1.1. Representing causal mechanisms

The use of mechanism information in everyday causal inference has been well documented. In causal attribution tasks (e.g., determining the cause of John's traffic accident), people overwhelmingly request information about causal mechanisms (e.g., asking whether John was drunk). That is, people attempt to fit the effect into their causal schemas or explanatory frameworks, using knowledge of generic causal mechanisms (here, alcohol consumption causes perceptual distortions, which can in turn cause traffic accidents) to link together the particular effect in question with its most likely cause (Ahn et al., 1995).

Mechanism information also influences performance in a variety of causal reasoning tasks. For example, causal mechanisms influence the use of causal vocabulary (e.g., "cause" and "prevent"; Walsh & Sloman, 2011). In common effect structures (i.e., *A* and *B* both are causes of *C*), the causes are seen as competing when they rely on distinct mechanisms, resulting in a discounting effect (Sloman, 1994; see also Waldmann & Holyoak, 1992 for related phenomena), but as mutually supporting when they rely on a common mechanism, resulting in a conjunction effect (Ahn & Bailenson, 1996). Moreover, in common cause structures (i.e., *A* is a cause of both *B* and *C*), people are more likely to violate the "screening off" principle (i.e., *B* and *C* are probabilistically independent, given knowledge of *A*) when the effects are seen as brought about by the same mechanism (Park & Sloman, 2013).

Although such ramifications of mechanism knowledge for causal judgment are extensive and well documented, less is known about how mechanism information is represented in semantic memory. One possibility is that people represent causal structures in interconnected webs or networks (see Fig. 1A), and when we recruit mechanism knowledge, we simply "zoom in" on a part of such a network. Alternatively, people may isolate and "chunk" (Chase & Simon, 1973; Miller, 1956) particular cause–effect configurations or mechanisms as individual units in memory, without also storing the connections between those mechanisms and other potentially contiguous mechanisms (see Fig. 1B). In what follows we explain each of these two views in detail.

The network approach has a long tradition in cognitive psychology. According to network or spreading activation theories of semantic memory (e.g., Anderson, 1983; Collins & Loftus, 1975), concepts are linked to one another via various kinds of associative links (such as "is a," "has a," etc). These theories have enjoyed considerable success in explaining a variety of memory effects. In particular, the number and strength of links to be traversed predict the likelihood of semantic priming, the degree of interference, and the time required to verify a semantic relationship (Anderson, 1983).

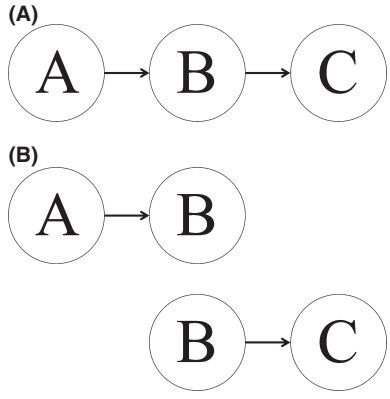

Fig. 1. (A) A network-based representation of a causal chain $A \rightarrow B \rightarrow C$. (B) A schema-based representation of a causal chain $A \rightarrow B \rightarrow C$, where A and B are chunked, B and C are chunked, but A and C are not chunked.

On the view that causal mechanisms are represented in a network-based manner, the various causal mechanisms that we have stored in memory would be linked together if they share at least one common variable. For example, the event type *Exercising* would be linked to *Becoming Thirsty*, which would be linked to *Drinking Water*. Similarly, *Having Sex* would be linked to *Becoming Pregnant*, which would be linked to *Becoming Nauseous.* Bayesian network theories of causation (e.g., Pearl, 2000; Spirtes, Glymour, & Scheines, 1993) would be a quintessential example of such an approach. For example, Glymour and Cheng (1998) provide the following example of a causal mechanism (from Baumrind, 1983):

> The number of never-married persons in certain British villages is highly inversely correlated with the number of field mice in the surrounding meadows. [Marriage] was considered an established cause of field mice by the village elders until the mechanisms of transmission were finally surmised: Never-married persons bring with them a disproportionate number of cats. (p. 295)

In this example, the number of cats is said to be a mechanism mediating the relationship between marriage and the number of mice. Representing this mechanism in a network would require a link from *Unmarried People* to *Cat Population*, and from *Cat Population* to *Mouse Population*. Because inferences traversing any number of links are in principle computable using Bayesian networks, a network-based representation would lend itself naturally to psychological versions of Bayes net theories (e.g., Gopnik et al., 2004; Griffiths & Tenenbaum, 2005). Although these theories would not require that all possible events be linked in a giant network, these theories *do* require that causally adjacent events should be represented in a locally connected manner—that is, as a causal network. The general success of these theories in modeling causal reasoning lends some initial plausibility to a network-based representation of causal mechanisms.

Network representations would predict that causation is generally *transitive*—that is, when *A* causes *B* and *B* causes *C*, then *A* would cause *C* (see Fig. 1A). For example, exercise (*A*) leads to thirst (*B*), and thirst (*B*) leads to drinking water (*C*), and intuitively, it also seems that exercise (*A*) leads to drinking water (*C*). Indeed, David Lewis (1973, 2000) argued that normatively, causation is *always* transitive (see also Strevens, 2008), and transitive inferences have been demonstrated empirically in studies using artificial stimuli (e.g., Ahn & Dennis, 2000; Goldvarg & Johnson-Laird, 2001; Von Sydow, Meder, & Hagmayer, 2009). Since network theories predict causal transitivity, these previous findings of causal transitivity weigh in favor of such theories (but see General Discussion).

Alternatively, causal mechanisms could be stored in relatively isolated *chunks* (Chase & Simon, 1973; Miller, 1956) or *schemas* (Alba & Hasher, 1983; Bartlett, 1932; Schank & Abelson, 1977). On such a view, some configurations of events might be stored as one coherent, schematized mechanism, such as *Exercise* causing *Thirst* causing *Water Consumption*. In other cases, even though two different mechanisms share an event in common, they may be stored separately. For instance, *Sex* causing *Pregnancy* would likely be stored as one mechanism, and *Pregnancy* causing *Nausea* as a separate mechanism, as illustrated in Fig. 1B.

Schematized causal mechanisms may be useful ways to organize clusters of events that co-occur in coherent and reliable causal patterns, just as concepts are useful ways of organizing clusters of correlated features (Rosch & Mervis, 1975; see also Zacks & Tversky, 2001 on clustering in event perception). Because we have concepts such as tigers, we can readily make inductive inferences such as "they are dangerous" (Murphy, 2002). Likewise, a causal mechanism for exercise can allow us to infer that a person would become thirsty after exercising.

At the same time, clustering events or features into concepts, schemas, or causal mechanisms necessarily entails discrete representations (see Dietrich & Markman, 2003; Markman, 1999). For instance, dogs and cats are disparate categories without overlap in category membership. This discretization can lead to striking categorical perception effects, wherein stimuli that are objectively very similar are perceived as psychologically much more distant because they belong to different categories or schemas (e.g., Eimas, Siqueland, Jusczyk, & Vigorito, 1971; Livingston, Andrews, & Harnad, 1998).

Our core prediction is that discrete representations of causal mechanisms can lead to causal intransitivity. That is, we propose that causal chains of the form $A \rightarrow B \rightarrow C$ are transitive only when *A*, *B*, and *C* are represented in the same schema. This prediction follows from the "narrative" strategy often used in everyday causal reasoning, wherein people reject a causal relationship between two events if one cannot use background knowledge to generate a "story" leading from the cause to the effect (e.g., Kahneman & Tversky, 1982; Taleb, 2007). If people represent mechanisms in terms of schemas rather than networks, then an inability to apply a schema leading from *A* to *C* would weaken the perceived causal strength of the $A \rightarrow C$ link. For example, *Exercise*, *Thirst*, and *Water Consumption* are likely to be stored in the same schema, so we would predict that since exercise (*A*) causes thirst (*B*) and thirst (*B*) causes water consumption (*C*), people would

therefore also think that exercise (*A*) causes water consumption (*C*). However, in other cases, *A* and *B* share a schema and *B* and *C* share a schema, but *A* and *C* do not share a schema, as in the sex/pregnancy/nausea example. Thus, even though sex (*A*) causes pregnancy (*B*) and pregnancy (*B*) causes nausea (*C*), people would be hesitant to agree that sex (*A*) causes nausea (*C*). As noted, a network representation, in linking causal schemas together that share variables in common, commits itself to even this second case being judged transitive. Any difference between the strength of the link between *Exercise* and *Water Consumption* and the link between *Sex* and *Nausea* would have to be due to the strength of the intermediate links, not due to the topology of the representation.

The schema and network theories therefore make quite different predictions about causal intransitivity, which we test in Experiments 1–3. These experiments focus on a set of 22 items, which were designed to vary in the extent to which *A*, *B*, and *C* formed a single schematized mechanism, or two disparate mechanisms that were joined by the *B* event. First, we demonstrate that there is considerable variety in schematization across these chains (Experiment 1). The exercise/thirst/drinking chain, for example, is highly schematized, whereas the sex/pregnancy/nausea chain is much less schematized. We also show that all these chains constitute mechanisms in the sense allowed by the network theory, that *B* always explains why *A* led to *C*, just as the number of cats explains why an increase in the number of singles led to a decrease in the number of mice (Glymour & Cheng, 1998). Next, we look at the consequences of schematization for causal transitivity. Whereas the network theory would predict that all causal chains should be transitive so long as the intermediate links ($A \rightarrow B$ and $B \rightarrow C$) are equally strong, the schema theory would predict that less schematized chains would be intransitive even if both links are very strong. We therefore obtain both the strength of the intermediate links and the transitivity of each causal chain, to see whether transitivity can vary even when the intermediate links are equally strong (Experiments 2 and 5). Then, we extend these results to a recognition memory paradigm, to provide converging evidence for causal intransitivity as well as to support our claim that these findings are consequences of the organization of memory (Experiment 3).

Note, however, that causal intransitivity can occur for reasons other than chunking. We address these alternative interpretations in Experiments 4 and 5. In the next section, we briefly review these additional sources of intransitivity.

1.2. Normative sources of intransitivity

Previous writers have identified a number of reasons why causal chains can be normatively intransitive (e.g., Björnsson, 2006; Broadbent, 2012; Hitchcock, 2001). A worrisome possibility is that we may find a relationship between schematization and transitivity, but for the opposite of the reason we are claiming. Whereas we are claiming that being unschematized leads a causal chain to be judged intransitive, it could instead be that some causal chains are normatively intransitive, leading them not to be schematized. Therefore, we must rule out these normative reasons for our causal chains to be intransitive. Based on Hitchcock (2001) and our own review of the literature, we found

five reasons why causality can be normatively intransitive. We call these *threshold effects* (Hausman, 1992), *incompatible aspects* (McDermott, 1995; Paul, 2000; Schaffer, 2005), *alternative causal pathways* (Eells & Sober, 1983), *petering out* (Lowe, 1980), and *lack of necessity or sufficiency* (e.g., Bonnefon, Da Silva Neves, Dubois, & Prade, 2008, 2012).

First, a *threshold effect* occurs when A influences the value of B, and B influences the value of C, but the influence of A on B does not push B beyond some threshold necessary to influence the value of C (Hausman, 1992). For example, swimming in the ocean causes the ingestion of salt water ($A \rightarrow B$), and the ingestion of salt water causes dehydration ($B \rightarrow C$). Yet a person who swims in the ocean will only ingest a small amount of salt water (say, 50 mL), and a larger amount (say, 200 mL) is required for dehydration to occur, so swimming in the ocean does not cause dehydration. One way of understanding threshold effects is that rather than $A \rightarrow B \rightarrow C$ constituting an intransitive causal chain, the underlying claims are in fact $A \rightarrow B_1$ and $B_2 \rightarrow C$, where $B_1 \neq B_2$ but B_1 and B_2 have identical linguistic descriptions because quantitative information is omitted (see also Lewis, 2000). Threshold effects are only possible when quantitative information is left tacit for B_1 and B_2, but the underlying values are different for B_1 and B_2—in the salt water case, B_1 is "ingesting 50 mL of salt water" and B_2 is "ingesting 200 mL of salt water," whereas both are described as "ingesting salt water."

Second, in cases of *incompatible aspects*, the property of the intermediate event B modified by A is not relevant to whether C occurs. Suppose a terrorist plans to detonate a bomb using his right hand, when his right hand is bit by a dog (A), causing him to push the button with his left hand (B), causing the bomb to explode (C); the dog bite did not cause the bomb to explode (McDermott, 1995). One account of this case (Paul, 2000; Schaffer, 2005) holds that causal claims are understood relative to the *contrast* they invoke—that is, "A caused B" means that A (rather than A') caused B (rather than B'). Depending on what aspects of A and B are singled out in this contrast, the truth of the causal claim could come out differently. In this case, the dog bite caused the terrorist to push the button with his left hand rather than with his right hand (call this contrast B_1), but the dog bite did *not* cause the terrorist to push the button with his left hand rather than not push the button at all (contrast B_2). Therefore, the aspect of B modified by A is not causally relevant to C—the dog bite causes B_1, but it is B_2 that causes the bomb to explode. Although B can be described as a left-handed button-pushing, it is this event *qua* left-handed (B_1) that is caused by A, and *qua* button-pushing (B_2) that causes C. This sort of intransitivity has been empirically documented in causal learning experiments where a dichotomous event A causes a change to one aspect of a complex category B, and a *different* aspect of B causes a change to a dichotomous event C (Hagmayer, Meder, von Sydow, & Waldmann, 2011). In such cases, aspects of the intermediate event that are qualitatively heterogeneous at the token level are described homogeneously at the type level (i.e., "A influences B" and "B influences C"). Transitivity is not normatively guaranteed under such circumstances.

Third, transitivity in causal chains can fail because of *alternative preventive causal pathways* from A to C (Eells & Sober, 1983; Hitchcock, 2001). For example, Nancy may

be dirty at time t_1, causing her to take a shower ($A \rightarrow B$), and taking a shower may cause her to be clean at t_2 ($B \rightarrow C$). But it seems odd to say that Nancy being dirty at t_1 caused her to be clean at t_2. Intransitivity occurs because there is also a preventive link between A and C; if one does not take a shower (i.e., holding constant the value of B), being dirty at t_1 makes it *less likely* that one will be clean at t_2. Such direct preventive pathways can result in violation of the Causal Markov Condition (see Pearl, 2000; Spirtes et al., 1993 and Experiment 4 below for details), making transitive inferences invalid.

Fourth, a causal chain can *peter out* over successive links (Lowe, 1980). In such cases, although each causal link is individually plausible, the causal link between the first and last event seems implausible. To borrow Lowe's (1980) example, consider the classic nursery rhyme:

For want of a nail the shoe was lost,
For want of a shoe the horse was lost,
For want of a horse the rider was lost,
For want of a rider the battle was lost,
For want of a battle the kingdom was lost,
And all for the want of a horseshoe nail.

While each causal link in this chain seems (at least somewhat) plausible, the overall causal connection between the first cause and the terminal effect is quite tenuous. According to some probabilistic analyses of causation (e.g., Jenkins & Ward, 1965), causal strength is proportional to the increase in the probability of the effect in the presence of the cause. If there are no alternative causal pathways between links, the causal strength between the first event in a causal chain "peters out" as the effects become more distant from the cause.

Finally, we consider the possibility that intransitivity may result from the relative necessity and sufficiency (e.g., Pearl, 2000) of A for B or B for C. Intuitively, some events seem to be causes because they are *sufficient* for an effect. For example, falling off a twenty-story building virtually guarantees death. Other relationships seem to be causal because they are *necessary*. For example, water is required for plants to grow. One possibility is that transitivity fails when intermediate causal strengths are low, as measured by the sufficiency or necessity of A for B or B for C. For example, sex is not a sufficient cause of pregnancy, so perhaps this is why the sex/pregnancy/nausea chain is intransitive. In particular, it has been proposed that the crucial factor for transitivity is whether A is a necessary cause for B in $A \rightarrow B \rightarrow C$ (Bonnefon et al., 2008, 2012; see Experiment 5 for details).

Throughout these experiments, we address the concern that some of our chains might be intransitive due to the normative factors mentioned above. Experiment 2 addresses the possibility that these findings are driven by threshold effects or by incompatible aspects, and Experiment 4 addresses concerns about alternative causal pathways, petering out, and lack of sufficiency or necessity. Finally, Experiment 5 tested for intransitivity in a new set of items that were specifically chosen to dissociate necessity and schematization.

Throughout these experiments, we predict that relatively unschematized causal chains will be intuitively transitive, even holding these normative factors constant.

2. Experiment 1

The goal of Experiment 1 was to empirically develop a set of causal chains varying in the extent to which they were schematized. We used 22 causal chains (Table 1), which we anticipated would vary in schematization. Each chain consisted of three temporally connected events (*A*, *B*, *C*) in which adjacent events (*A* and *B*, or *B* and *C*) would likely be perceived as causally connected (see Experiment 2 for empirical support). In Experiment 1, we measured two different senses in which these chains could comprise causal mechanisms.

First, we measured the extent to which each causal chain was schematized, or represented as one coherent unit in semantic memory. To measure this, participants were asked to rate the extent to which *B* needed to be explicitly mentioned to explain to another person why *A* led to *C*. We asked for judgments about explaining the causal chain to *others* to reduce the possibility of hindsight bias (e.g., believing that one could have easily inferred pregnancy after hearing that sex led to nausea), relying on participants' assumption that schemas would be culturally shared common knowledge. To the extent that a causal chain is schematized, it should be possible for most people to understand the causal chain without explicitly mentioning the intermediate event *B*, because *B* is represented in the same schema as *A* and *C* and would be inferred automatically. This follows from the critical role of schemas in inductive inference (Bartlett, 1932; Schank & Abelson, 1977). For example, upon hearing that Allison exercised and drank water, one would infer automatically that she became thirsty because all three of these events belong to a common schema. A speaker would therefore be unlikely to mention thirst in explaining why Allison's exercising led her to drink. But to the extent that a causal chain is not schematized, explicit mention of *B* would be necessary to understand the relationship between *A* and *C*, since *A* and *C* do not share a common schema, but are only related to one another in virtue of *B*. Upon hearing that Francine had sex and experienced nausea, the inference that her nausea was caused by pregnancy is effortful or sometimes impossible, and it would be infelicitous for a speaker to omit this fact. We predicted that our items would vary considerably in the perceived need to mention *B*, as the items were designed to vary in the extent to which they had underlying schematized mechanisms.

Second, as mentioned earlier, some writers have suggested that what it means to be a causal mechanism is to fully mediate or explain a causal relationship (e.g., Glymour & Cheng, 1998). We measured this *explanatory* sense of mechanism by asking participants to rate the extent to which *B* explains why *A* led to *C*. If people have network representations of causal mechanisms, then intermediate links can potentially differ in strength, leading to variation in the extent to which causal chains are mechanisms in this explanatory sense. In developing stimuli, we aimed to equate the extent to which our chains constituted causal mechanisms in this explanatory sense relevant to the network view.

Table 1
Stimuli and results from Experiments 1 and 2

Item	Event A	Event B	Event C	Experiment 1		Experiment 2		
				Explanation	Chunking	$A \rightarrow B$	$B \rightarrow C$	$A \rightarrow C$
1	Allison exercised for 20 min, then	Allison became thirsty, then	Allison drank a whole bottle of water.	8.18	7.46	8.17	8.40	8.07
2	Zach got distracted while grilling, then	Zach left the steak on the grill a bit too long, then	Zach's steak got scorched.	8.64	7.30	8.13	8.60	8.17
3	Wanda forgot to put on sunblock, then	Wanda's skin was hit by UV rays, then	Wanda's skin turned red.	8.43	7.22	6.03	8.60	7.60
4	Pam did not floss her teeth every day, then	Pam's teeth often had plaque on them, then	Pam developed cavities.	7.75	7.12	8.13	8.33	8.03
5	Melissa was outside in warm weather, then	Melissa's body temperature rose, then	Melissa's clothes were soaked with sweat.	7.79	7.09	8.17	8.60	7.73
6	Larry's alarm did not go off, then	Larry overslept by 10 min, then	Larry was late to work.	8.14	7.07	8.43	8.37	8.13
7	Carl studied for a while, then	Carl learned most of the material, then	Carl got a perfect score on the test.	7.25	6.74	8.47	8.57	8.27
8	Yarron was assigned to too many tasks, then	Yarron felt very stressed, then	Yarron was unable to concentrate.	7.64	5.04	8.30	8.20	7.67
9	Tess ate rancid pork, then	Tess developed a minor fever, then	Tess went to the doctor.	7.57	4.83	7.93	8.23	8.00
10	Erica did too much yoga, then	Erica strained her muscles, then	Erica went to get a massage.	7.89	4.35	8.17	7.97	6.67
11	Brad had many glasses of wine, then	Brad fell asleep, then	Brad had a dream.	6.61	3.85	7.33	8.13	4.10
12	Oscar's computer was infected by a virus, then	Oscar lost some of his computer files, then	Oscar's work was delayed by a week.	8.18	3.81	8.40	8.03	8.30
13	The dog was aging, then	The dog died, then	The dog was buried.	8.57	3.81	8.07	8.70	5.63
14	The wind blew, then	Ivan's hat was blown away, then	Ivan had to look for a new hat.	8.68	3.74	8.47	8.07	6.13

(continued)

Table 1. (continued)

Item	Event A	Event B	Event C	Experiment 1		Experiment 2		
				Explanation	Chunking	$A \rightarrow B$	$B \rightarrow C$	$A \rightarrow C$
15	Francine had unprotected sex while ovulating, then	Francine became pregnant, then	Francine experienced nausea.	8.36	3.15	8.63	8.23	5.50
16	Stacy picked up a hot stew pot without potholders, then	Stacy dropped the stew pot, then	Stacy cleaned the floor.	8.32	2.42	8.40	8.23	6.00
17	Quincy bought an air conditioner, then	Quincy's electric bill went up, then	Quincy was mad at the electric company.	7.64	2.21	7.83	8.10	5.27
18	Rhea wore high heels for a long time, then	Rhea developed back pain, then	Rhea missed work.	7.39	2.08	7.53	7.23	6.00
19	George was late for a meeting, then	George drove very fast, then	George annoyed other drivers.	7.32	1.83	8.00	8.20	6.67
20	Xavier played with Legos, then	Xavier's room became messy, then	Xavier's mother became upset	8.11	1.81	7.30	7.83	3.73
21	Karen stepped on a dog, then	The dog growled loudly, then	A child was scared.	8.25	1.77	8.43	8.17	5.73
22	Ned ate very spicy food, then	Ned drank a lot of water, then	Ned had to urinate.	8.21	1.73	8.00	8.57	4.73

Note. Chunking scores were calculated by reverse-coding the "need to mention" scores from Experiment 1.

Thus, we anticipated that all 22 of our chains would comprise causal mechanisms in this sense (i.e., *B* explaining why *A* led to *C*) because we selected all these items such that *A* strongly causes *B* and *B* strongly causes *C* (see Experiment 2 for empirical support). That is, this definition of mechanism should apply not only to the chains we would expect to be highly schematized (e.g., exercise, thirst, and drinking water) but also to the chains we would expect to be relatively unschematized (e.g., sex, pregnancy, and nausea). Consequently, any subsequent differences in inferences that we find between more and less schematized causal chains would be difficult to explain on the network view.

2.1. Methods

Participants in all experiments were recruited and compensated through Amazon Mechanical Turk and were from the United States. Measures were taken to prevent participants from completing multiple experiments reported in this article. Thirty participants were recruited for Experiment 1 and two were excluded for providing random ratings for noncausal filler chains (see below).

Materials include 22 test items and 11 filler items. The test items were sets of three events that can form a causal chain ($A \rightarrow B \rightarrow C$) where we expected both intermediate links to be highly causal (see Table 1). The filler chains contained one or more noncausal links and were used to detect random responding (e.g., Hannah ate a hash brown, then Hannah wore a red hat, then Hannah won the lottery). For every chain, participants first completed the *Explanation* measure:

> Consider the following events, *A* and *C*:
> *A*: Carl studied for a while, then
> *C*: Carl got a perfect score on the test
> Now consider the following event, *B*, which occurred between *A* and *C* as follows:
> *A*: Carl studied for a while, then
> ***B*: Carl learned most of the material**, then
> *C*: Carl got a perfect score on the test

Participants then rated "To what extent do you think that event *B* explains why *A* led to *C*?" on a scale from 1 ("B does not at all explain why A led to C") to 9 ("B fully explains why A led to C"). Because we anticipated that these ratings would be at ceiling for the test chains but at floor for the noncausal filler chains, we excluded from analysis two participants whose scores on the noncausal filler items were more than two *SDs* above the mean on this measure.

If a participant's rating on the *Explanation* measure was above the scale midpoint (5) for a given item, they completed the *Need to mention* measure on the following screen. They were told to "consider again the events *A*, *B*, and *C*" from the previous screen, which were listed as a reminder. They then were asked to "suppose you were explaining to someone how event *A* led to event *C*" and rated "To what extent do you think that it is essential to explicitly mention *B* in explaining how *A* led to *C*?" on a scale from 1 ("B

is not important to mention because it is obvious") to 9 ("B is important to mention because it is not obvious"). This question was only asked for participants who had agreed that *B* explained why *A* led to *C*, since the *Need to mention* question is ill-defined if *B* does not explain the relationship between *A* and *C*. For instance, it would be unnecessary to mention Hannah's wearing a red hat (*B*) in explaining how Hannah's eating a hash brown (*A*) led to Hannah's winning the lottery (*C*). However, this judgment does not reflect a belief that the three events are schematized, but rather a belief that *B* does not even explain why *A* led to *C*.

The 33 chains were presented in a random order. All experiments reported in this article were conducted online using Qualtrics software, with no time limit except as indicated.

2.2. Results and discussion

As anticipated, *B* was not judged to explain why *A* led to *C* for the filler items ($M = 1.27$, $SD = 0.17$), but was judged highly relevant in explaining why *A* led to *C* for the test items ($M = 7.95$, $SD = 0.52$). In addition, all test items were rated significantly above the scale midpoint (all ps $< .05$ using the False Discovery Rate procedure to correct for multiple comparisons; Benjamini & Hochberg, 1995). Therefore, for all 22 of our test items, *B* constituted a causal mechanism in the sense of mediating the relationship between *A* and *C*. The mean *Explanation* ratings for each test item are shown in Table 1.

To evaluate the extent to which participants believed that the causal chains were chunked or schematized into single causal mechanisms, we reverse-coded the *Need to mention* question to form a measure we will refer to as *Chunking*. That is, for chains where it was essential to mention *B*, this was indicative of low chunking, and for chains where it was not essential to mention *B*, this was indicative of high chunking. The mean Chunking ratings for each test item are shown in Table 1. Among our set of test items, the explanation measure had no association with the chunking ratings, $r(20) = .08$, $p = .74$.

These results show that factors other than the extent to which *B* explains why *A* causes *C* can influence the extent to which causal mechanisms are represented as single chunks in long-term memory. This finding is not straightforwardly explained by a network-based theory of causal representation but is consistent with a schema-based theory, which allows causal chains to be discretely represented in multiple disparate schemas even if they share a common variable. It may nonetheless be perfectly clear that *B* mediates the relationship between *A* and *C* in the case at hand—as acknowledged by our participants. In Experiments 2 and 3, we will explore the consequences of schematization for the transitivity of causal chains.

3. Experiment 2

As we suggested in the introduction, causal chains that are chunked together seem to be *transitive*—when *A* causes *B* and *B* causes *C*, *A* also seems to cause *C*—yet this does not

seem to be true for the chains that consist of two disparate mechanisms. In Experiment 2, we sought to verify this claim. In particular, we predicted that the more a chain was shown to be schematized in Experiment 1, the more transitive it would be judged to be. Thus, the relatively unschematized chains would be seen as less transitive—that A would cause B and B would cause C, but A would not cause C. We also measured the causal strength of the intermediate ($A \rightarrow B$ and $B \rightarrow C$) links to ensure that intransitivity did not occur merely because the more schematized chains had stronger intermediate links.

3.1. Methods

Thirty-two participants were recruited for Experiment 2 and two were excluded for providing random ratings for the noncausal filler chains. For each of the 22 test chains and 11 filler chains from Experiment 1, participants provided ratings of $A \rightarrow B$, $B \rightarrow C$, and $A \rightarrow C$ causality as follows. For each chain, participants first saw the three events (e.g., "Carl studied for a while, then Carl learned most of the material, then Carl got a perfect score on the test") and were asked, "To what extent would you say that: [X] caused [Y]" where X and Y were filled in with A and B, B and C, or A and C from Table 1 (e.g., "Carl studying for a while caused Carl to learn most of the material"). Ratings were provided on a 9-point scale (1: "definitely would not"; 5: "unsure"; 9: "definitely would"). For each chain, ratings were elicited in the same fixed order ($A \rightarrow B$, $B \rightarrow C$, $A \rightarrow C$). This order creates some demand to infer $A \rightarrow C$ after responding affirmatively to $A \rightarrow B$ and $B \rightarrow C$, providing a stronger test against our hypothesis that some chains are intransitive. Each chain was presented on a separate screen in a random order.

3.2. Results and discussion

As shown in Table 1, the intermediate ($A \rightarrow B$ and $B \rightarrow C$) links were rated very strong for virtually all the test chains ($M = 8.02$, $SD = 0.57$ for $A \rightarrow B$; $M = 8.24$, $SD = 0.33$ for $B \rightarrow C$), with only one intermediate link (of 44) not significantly higher than the scale midpoint (i.e., $p < .05$ using the False Discovery Rate procedure to correct for multiple comparisons; Benjamini & Hochberg, 1995). In contrast, the strength judgments of the $A \rightarrow C$ links were much lower overall ($M = 6.64$) and much more variable ($SD = 1.44$). Thus, these chains vary in the extent to which they are seen as transitive or intransitive, even though all intermediate links are very strong.

To test the hypothesis that intransitivity occurs when causal mechanisms are not chunked together into one schema, we used the chunking ratings from Experiment 1 to predict $A \rightarrow C$ causal ratings in a multiple regression, using item means as the unit of analysis. As hypothesized, chunking significantly predicted $A \rightarrow C$ causal ratings, $b = 0.52$, $SE = 0.09$, $p < .001$. However, it is also critical to ensure that that the $A \rightarrow C$ ratings are not lower for some chains simply because the $A \rightarrow B$ and $B \rightarrow C$ links are weaker (i.e., a "petering out" effect at the token level; Lowe, 1980). To verify that differences in the strength of the intermediate links were not responsible for our intransitive

chains, we conducted another regression predicting $A \rightarrow C$ causal ratings from chunking, but also including $A \rightarrow B$ and $B \rightarrow C$ causal strength as adjustment variables. The effect of $A \rightarrow B$ causal strength reached marginal significance, $b = 0.68$, $SE = 0.34$, $p = .060$, whereas the effect of $B \rightarrow C$ causal strength was not significant, $b = -0.67$, $SE = 0.70$, $p = .35$. Most important, the effect of chunking remains significant when adjusting for these other variables, $b = 0.58$, $SE = 0.11$, $p < .001$.[1] Because the $A \rightarrow C$ links differ in strength despite equally strong $A \rightarrow B$ and $B \rightarrow C$ links, a network theory of causal representation cannot straightforwardly explain this result.

Could there be some normative reason why some of these causal chains were judged to be intransitive? In the introduction, we outlined five reasons why a causal chain can be normatively intransitive—threshold effects, incompatible aspects, alternative causal pathways, petering out, and lack of necessity or sufficiency. If a causal chain is perceived as intransitive for one of these reasons, this could potentially undermine our claim that it is intransitive due to its failure to be schematized—instead, that chain could fail to be schematized simply because it is normatively intransitive. We defer consideration of three of these accounts—alternative preventive pathways, petering out at the type level, and lack of necessity or sufficiency—until Experiment 4. However, the design of this experiment speaks against threshold effects or incompatible aspects as explanations of our intransitive chains.

A threshold effect occurs when A causes B to have a certain value (say, B_1), and B only affects C if B is set above a threshold (say, B_2) that is higher than B_1 (Hausman, 1992). To borrow the example from the introduction, swimming in the ocean causes the ingestion of a small amount of salt water (say, 50 mL), and the ingestion of a larger amount (say, 200 mL) causes dehydration, so swimming in the ocean does not cause dehydration. In this experiment, however, participants were presented with A, B, and C at the token level, and B was the same token event in both $A \rightarrow B$ and $B \rightarrow C$. As a result, B_1 and B_2 would have been equated, so a threshold effect cannot have occurred.

Likewise, these results are unlikely to be due to incompatible aspects of the intermediate event (Paul, 2000; Schaffer, 2005). In such cases, intransitivity occurs because the property of B affected by A is different from the property of B that is responsible for affecting C. Yet, in Experiment 1, the intransitive chains contained intermediate events in which only one property could plausibly be involved in either the $A \rightarrow B$ or $B \rightarrow C$ relation. For example, drinking wine caused one to sleep (rather than not sleep) and sleeping (rather than not sleeping) caused one to have a dream. To make the strongest case against this account, consider a relatively intransitive chain that seems to be the most plausible candidate for such an explanation: "Karen stepped on a dog, then the dog growled loudly, then a child was scared" (item 21 in Table 1). Perhaps the causal relations should be decomposed as "Karen stepped on a dog, causing the dog to growl loudly (rather than not growl at all)" and "The dog growled loudly (rather than quietly), causing a child to be scared." Yet the $B \rightarrow C$ relation cannot depend on the loudness of the growl rather than the growl itself, because "the dog growled quietly, causing the child to be scared" is a perfectly reasonable claim. Similar linguistic tests seem to rule out all plausible alternative contrasts for the other intransitive chains, rendering it unlikely that incompatible aspects can explain our intransitive causal chains.

In the next experiment, we seek converging evidence for the relationship between schematization and transitivity, using a different measure of transitivity.

4. Experiment 3

People falsely recognize new sentences that are implied by previously memorized sentences (e.g., Bransford, Barclay, & Franks, 1972). For example, given, "The hungry python caught the mouse," more participants falsely recalled, "The hungry python ate the mouse" than correctly recalled the original sentence (Brewer, 1977). Experiment 3 used this paradigm to test transitive inference from causal sentences. After memorizing the $A \rightarrow B$ sentences (e.g., "Carl studying for a while caused Carl to learn most of the material") and $B \rightarrow C$ sentences ("Carl learning most of the material caused Carl to get a perfect score on the test"), participants' recognition memory was tested using the old sentences plus $A \rightarrow C$ ("Carl studying for a while caused Carl to get a perfect score on the test") and $C \rightarrow A$ ("Carl getting a perfect score on the test caused Carl to study for a while") as new sentences. If participants automatically infer $A \rightarrow C$ from $A \rightarrow B$ and $B \rightarrow C$, they would falsely recognize the $A \rightarrow C$ sentences but not the $C \rightarrow A$ sentences. Thus, while the false alarm rate for $C \rightarrow A$ sentences would be low for all chains, the false alarm rate for $A \rightarrow C$ sentences would be high to the extent that a chain is transitive, because $A \rightarrow B$ and $B \rightarrow C$ imply $A \rightarrow C$ to the extent that a causal chain is transitive. We therefore anticipated more false recognition of the $A \rightarrow C$ sentences for the more schematized chains.

4.1. Methods

Thirty-six participants were recruited and six were excluded because their performance was at chance. Participants were instructed that they were participating in a memory experiment where they were to identify new and old sentences. Participants then completed four practice items, which involved temporal but not causal relations, such as "John bought sunglasses, then John went to the beach" (*A–B*) and "John went to the beach, then John made a sandcastle" (*B–C*). Just as in the recognition task in the main experiment, they responded to those old sentences, as well as new sentences of the form "John bought sunglasses, then John made a sandcastle" (*A–C*) and "John made a sandcastle, then John bought sunglasses" (*C–A*). Accuracy feedback was provided to ensure that any findings from the main task were not due to participants' misunderstanding the nature of the task and believing that logical implications count as old sentences.

Next, participants saw the 22 test chains from Table 1, in a random order. Each screen contained the two old sentences for each chain, with $A \rightarrow B$ above $B \rightarrow C$. Participants could look at each item for up to 10 s or press the space bar to advance to the next item when ready. Because the old sentences included $A \rightarrow B$ and $B \rightarrow C$ for each of the 22 test chains used in Experiments 1 and 2, there were 44 old sentences in total.

Immediately afterward, participants completed the recognition task. They were told that they would see sentences of which half were old and half were new, and that their task was to classify the sentences in fewer than 5 s, by pressing "y" for an old sentence and "n" for a new sentence. They were instructed explicitly that a sentence did not count as old unless it was presented word for word, and that the task was not to identify logical inferences. All 44 sentences used in the presentation phase were used as old sentences. The new sentences were $A \rightarrow C$ and $C \rightarrow A$ for each of the 22 test chains, and there were thus 44 new sentences. Each sentence was presented on a separate screen in a random order. No accuracy feedback was provided.

4.2. Results and discussion

For our main analyses, we examined the proportion of participants committing a miss (i.e., responded "no" to an old item) or a false alarm (i.e., responded "yes" to a new item) for each item (see Table 2). The proportion of misses was low and did not significantly differ between $A \rightarrow B$ sentences ($M = 0.11$, $SD = 0.08$) and $B \rightarrow C$ sentences

Table 2
Proportion of errors for each item in Experiment 3

Item	$A \rightarrow B$	$B \rightarrow C$	$A \rightarrow C$	$C \rightarrow A$
1	0.13	0.10	0.57	0.07
2	0.15	0.17	0.63	0.03
3	0.17	0.03	0.56	0.10
4	0.07	0.23	0.34	0.03
5	0.13	0.17	0.45	0.07
6	0.13	0.23	0.63	0.13
7	0.21	0.20	0.50	0.00
8	0.10	0.14	0.52	0.07
9	0.03	0.23	0.40	0.07
10	0.03	0.18	0.20	0.03
11	0.07	0.17	0.37	0.03
12	0.07	0.24	0.59	0.07
13	0.14	0.17	0.20	0.07
14	0.10	0.07	0.27	0.07
15	0.07	0.17	0.20	0.00
16	0.00	0.17	0.14	0.07
17	0.13	0.17	0.10	0.03
18	0.07	0.28	0.13	0.00
19	0.33	0.03	0.14	0.07
20	0.03	0.17	0.17	0.00
21	0.07	0.13	0.17	0.00
22	0.27	0.07	0.13	0.00

Note. Items are listed in descending order of chunking scores from Experiment 1 (that is, the same order as Table 1). The $A \rightarrow B$ and $B \rightarrow C$ columns give miss proportions, and the $A \rightarrow C$ and $C \rightarrow A$ columns give false alarm proportions.

($M = 0.16$, $SD = 0.07$), $t(21) = -1.67$, $p = .11$, $d = 0.36$. However, false alarms were much more frequent for $A \rightarrow C$ ($M = 0.34$, $SD = 0.19$) than for $C \rightarrow A$ sentences ($M = 0.05$, $SD = 0.04$), $t(21) = 7.90$, $p < .001$, $d = 1.68$.

Most critically, the proportion of false alarms for $A \rightarrow C$ sentences for each item was very strongly correlated with that item's chunking score from Experiment 1, $r(20) = .85$, $p < .001$. The miss proportions were not correlated with chunking for either $A \rightarrow B$, $r(20) = .07$, $p = .76$, or for $B \rightarrow C$, $r(20) = .12$, $p = .59$, indicating that memory was not simply worse overall for the items with higher chunking scores. Instead, it appears that for the more chunked items, participants represent A and C as parts of the same causal mechanism, leading to a high false recognition rate for the $A \rightarrow C$ sentences. Consistent with this possibility, the $C \rightarrow A$ false alarm proportion, though low overall ($M = 0.05$), was somewhat higher for the more chunked items, $r(20) = .46$, $p = .030$, suggestive of a stronger association between A and C for the more chunked items.

Although petering out at the token level could not explain the intransitivity found in Experiment 2, we conducted a follow-up analysis to verify that the effect of chunking on $A \rightarrow C$ false alarms could also withstand adjustment for $A \rightarrow B$ and $B \rightarrow C$ causal strength. A multiple regression predicting each item's proportion of $A \rightarrow C$ false alarms from chunking (Experiment 1) and $A \rightarrow B$ and $B \rightarrow C$ causal strength ratings (Experiment 2) revealed that only chunking was a significant predictor of $A \rightarrow C$ false alarms, $b = 0.08$, $SE = 0.01$, $p < .001$ (for $A \rightarrow B$ causal ratings, $b = 0.01$, $SE = 0.04$, $p = .86$; for $B \rightarrow C$ causal ratings $b = -0.04$, $SE = 0.09$, $p = .68$). Once again, petering out cannot explain the variation in transitivity among items, but the chunking of causal mechanisms can.

These findings add to the results of Experiment 2 in showing that chunking predicts transitivity, as measured by a very different kind of dependent variable—error rates on a recognition memory task. The false alarm rates for $A \rightarrow C$ sentences were *higher* for chains that were highly chunked, consistent with our claim that A and C were stored in the same schema for the more highly chunked chains.

5. Experiment 4

Although Experiment 2 rendered petering out at the token level, threshold effects, and incompatible aspects unlikely as explanations for our intransitive chains, three other alternative accounts remain on the table: alternative causal pathways, petering out at the type level, and lack of necessity and sufficiency. Experiment 4 examined whether the relationship between schematization and intransitivity holds up even after adjusting simultaneously for alternative causal pathways, probabilistic strength of the intermediate links, necessity, and sufficiency. Across Experiment 4A–D, we collected 12 conditional probability judgments relevant for those remaining three alternative accounts (see Table 3 for sample wordings). Then, we used multiple regression to predict the effect of chunking on transitivity, while adjusting for the potential effects of each alternative account.

In Experiment 4A, we examined the possibility that *alternative causal pathways* from A to C can explain the intransitive chains (Eells & Sober, 1983; Hitchcock, 2001). For

Table 3
Sample question wordings from Experiment 4

Experiment	Judgment	Example Question
4A	P(*C*|*A*,*B*)	Of 100 people who studied for a while and learned most of the material, how many would get a perfect score on the test?
	P(*C*|~*A*,*B*)	Of 100 people who did not study for a while but learned most of the material, how many would get a perfect score on the test?
	P(*C*|*A*,~*B*)	Of 100 people who studied for a while but did not learn most of the material, how many would get a perfect score on the test?
	P(*C*|~*A*,~*B*)	Of 100 people who did not study for a while and did not learn most of the material, how many would get a perfect score on the test?
4B	P(*B*|*A*)	Of 100 people who studied for a while, how many would learn most of the material?
	P(*B*|~*A*)	Of 100 people who did not study for a while, how many would learn most of the material?
	P(*C*|*B*)	Of 100 people who learned most of the material, how many would get a perfect score on the test?
	P(*C*|~*B*)	Of 100 people who did not learn most of the material, how many would get a perfect score on the test?
4C	Sufficiency of *A* for *B*	Consider 100 cases in which someone studies for a while. In how many cases will this cause them to learn most of the material?
	Sufficiency of *B* for *C*	Consider 100 cases in which someone learns most of the material. In how many cases will this cause them to get a perfect score on the test?
4D	Necessity of *A* for *B*	Consider 100 cases in which someone learns most of the material. In how many cases was this caused by their studying for a while?
	Necessity of *B* for *C*	Consider 100 cases in which someone gets a perfect score on the test. In how many cases was this caused by their learning most of the material?

Note. Wordings correspond to item 7 in Table 1, where *A* is "studying or a while," *B* is "learning most of the material," and *C* is "getting a perfect score on the test".

example, although drinking wine (*A*) causes sleep (*B*), and sleep causes dreaming (*C*), people may believe that wine-drinking prevents one from having a dream. If people believe in such alternative preventive causal pathways, apparent intransitivity (refusing to endorse that *A* causes *C*) may be simply due to this other, preventive path between *A* and *C*.

To test whether alternative preventive paths between *A* and *C* exist for our intransitive chains, we tested the "screening off" property for each chain (also known as the Causal Markov Condition; Pearl, 2000; Spirtes et al., 1993). If there is no alternative pathway between *A* and *C*, then *B* will "screen off" the influence of *A* on *C*, so that *A* provides no additional information about *C* once *B* is accounted for—that is, P(*C*|*A*,*B*) = P(*C*|~*A*,*B*) and P(*C*|*A*,~*B*) = P(*C*|~*A*,~*B*). If there is an alternative preventive pathway from *A* to *C* (e.g., if wine prevents dreaming), then screening off is violated: Even given that a person falls asleep, he or she is less likely to dream if he or she drank wine, so P(*C*|*A*,*B*) < P(*C*|~*A*,*B*). Put differently, an alternative preventive pathway from *A* to *C* means that the contingency between *A* and *C* would be negative when holding *B* constant. In Experiment 4A, one group of participants made frequency estimations corresponding to P(*C*|*A*,*B*) and

P(C|~A,B), allowing us to compute the A–C contingency when B is present [$\Delta P_{AC|B}$ = P(C|A,B) – P(C|~A,B)], and another group of participants made frequency estimates of P(C|A,~B) and P(C|~A,~B), allowing us to compute the A–C contingency when B is absent [$\Delta P_{AC|\sim B}$ = P(C|A,~B) – P(C|~A,~B)]. If either $\Delta P_{AC|B}$ or $\Delta P_{AC|\sim B}$ is negative for a chain, an alternative preventive pathway from A to C could lead that chain to be intransitive.

In Experiment 4B, we examined whether *petering out* at the type level could explain our findings (Lowe, 1980). Petering out occurs when the $A \rightarrow B$ and $B \rightarrow C$ links are viewed as weakly causal, but the strength of these intermediate links "peters out" so that the $A \rightarrow C$ link is not viewed as causal. Specifically, if the contingency between A and B, ΔP_{AB} = P(B|A) – P(B|~A), or the contingency between B and C, ΔP_{BC} = P(C|B) – P(C|~B), were low for a chain, then the contingency between A and C would be even lower.

Although participants in Experiment 2 judged the $A \rightarrow B$ and $B \rightarrow C$ links to be equally causal for the transitive and intransitive chains, these judgments were obtained at the token (particular) level rather than the type (category) level. Some causal relationships hold at the token but not the type level—for example, hitting a golf ball into a tree may cause a golfer to make a hole in one on some particular occasion, but one would not say that *in general* hitting golf balls into trees causes holes in one (Rosen, 1978). Depending on the strategy used for making the $A \rightarrow C$ causal judgments in Experiment 2, the strength of the type-level links could play a role. In particular, participants could have relied on temporal contiguity to answer the questions about the $A \rightarrow B$ and $B \rightarrow C$ links (Lagnado & Sloman, 2006), since the events occurred in a fixed order, but relied on the contingency of the intermediate links to answer the questions about the $A \rightarrow C$ links. If so, petering out at the type level could explain why participants sometimes gave low ratings for $A \rightarrow C$, even as temporal contiguity led them to endorse the $A \rightarrow B$ and $B \rightarrow C$ links. To assess this possibility, Experiment 4B measured P(B|A), P(B|~A), P(C|B), and P(C|~B) to compute ΔP_{AB} and ΔP_{BC}.

Finally, we measured the sufficiency and necessity of the intermediate links (Einhorn & Hogarth, 1986; Mackie, 1965; Pearl, 2000). An event A is sufficient for an event B if B always occurs when A occurs. For instance, falling off a 20-story building is a sufficient cause for death. In Cheng (1997), causal power is defined as the sufficiency of the cause for bringing about the effect. Experiment 4C measured the sufficiency of the intermediate links to test whether differences in intermediate strength as measured in terms of sufficiency can account for differences in transitivity.

Alternatively, necessity could play a role in transitivity. A is necessary for B if B can only occur when A occurs. For instance, watering a houseplant is necessary for it to grow. Bonnefon et al. (2008, 2012) propose that people often adopt a conception of causality on which causal chains are transitive when a *saliency condition* is met, such that A is such a necessary cause of B that observing B leads one to expect A. According to Bonnefon et al. (2008), we often equate actions (A) and their consequences (B) in causality ascriptions when the consequences are highly diagnostic of the actions, but not when the consequences are not highly diagnostic; in only the former case would transitivity be justified. Borrowing their example, suppose Cindy drives to the countryside (A), causing her license plate to become muddy (B), causing her to get a fine (C). In the case where A is

a salient or necessary cause of B—that is, Cindy's license plate getting muddy occurs only after she goes to the countryside—it seems appropriate to identify A with B, and to say that driving to the countryside caused Cindy to get fined. But, in the case where A is not necessary for B—say, because Cindy sometimes drives down her muddy driveway—it does not seem appropriate to say that going to the countryside caused her to get fined. Therefore, a lack of necessity in the $A \rightarrow B$ link could account for the intransitive chains. Experiment 4D tested this possibility.

5.1. Methods

We recruited 159 participants for Experiment 4 (n = 40, 39, 39, and 41 for Experiments 4A–D, respectively). For all four experiments, participants provided frequency estimates (Buehner, Cheng, & Clifford, 2003; Gigerenzer & Hoffrage, 1995) corresponding to the questions summarized in Table 3, for the 22 sets of test chains from Experiments 1–3. Before completing these ratings, they were instructed that "these questions will ask you to think about a sample of 100 people" and that this sample should be thought of as "a typical, representative sample of people in the United States." All ratings were completed on a sliding scale from 0 to 100.

For Experiment 4A, 18 participants estimated P(C|A,B) and P(C|~A,B) for each chain, and 22 participants estimated P(C|A,~B) and P(C|~A,~B) for each chain. Estimates of each type of probability judgment were blocked, so that a participant might have rated P(C|A,B) for all 22 chains (in a random order), followed by P(C|~A,B) for all 22 chains (in a different random order). The order of the blocks was counterbalanced. These scores were converted to $\Delta P_{AC|B}$ and $\Delta P_{AC|\sim B}$ by taking the difference between the two judgments (i.e., $\Delta P_{AC|B}$ = P(C|A,B) – P(C|~A,B) and $\Delta P_{AC|\sim B}$ = P(C|A,~B) – P(C|~A,~B)), which could potentially range from −100% to 100%.

For Experiment 4B, 18 participants estimated P(B|A) and P(B|~A) and 21 estimated P(C|B) and P(C|~B). These probability judgments were blocked, with the order of the 22 items randomized within each block. The positive block (e.g., P(B|A)) always preceded the negative block (e.g., P(B|~A)) to provide context for the negative judgment. These scores were converted to ΔP_{AB} and ΔP_{BC} by taking the difference between the two judgments (i.e., ΔP_{AB} = P(B|A) – P(B|~A) and ΔP_{BC} = P(C|B) – P(C|~B)), which could potentially range from −100% to 100%.

For Experiment 4C, 19 participants rated the sufficiency of the $A \rightarrow B$ link for each chain, and 20 participants rated the sufficiency of the $B \rightarrow C$ link for each chain, and for Experiment 4D, 21 participants rated the necessity of the $A \rightarrow B$ link for each chain, and 20 participants rated the necessity of the $B \rightarrow C$ link for each chain. These ratings were completed in a random order and could potentially range from 0% to 100%.

5.2. Results and discussion

We used multiple regression to test whether chunking continued to predict transitivity after adjusting for the effects of alternative causal pathways, petering out, and necessity

Table 4
Probability judgments for each item in Experiment 4

	Experiment 4A: Alternative Pathways		Experiment 4B: Contingency Ratings		Experiment 4C: Sufficiency Ratings		Experiment 4D: Necessity Ratings	
Item	$\Delta P_{AC\backslash \sim B}$	$\Delta P_{AC\backslash B}$	ΔP_{AB}	ΔP_{BC}	$A \rightarrow B$	$B \rightarrow C$	$A \rightarrow B$	$B \rightarrow C$
1	7.32	14.44	36.00	31.81	62.26	43.35	25.57	90.95
2	2.82	11.50	29.67	31.57	61.42	43.40	71.19	87.90
3	−3.50	23.17	13.83	30.43	87.84	43.35	58.19	57.85
4	10.05	−5.72	27.56	33.14	66.74	68.30	70.29	75.45
5	7.91	13.83	58.17	22.62	53.68	55.65	46.33	79.35
6	0.91	1.78	34.00	13.14	61.21	36.25	54.95	33.20
7	1.82	8.56	36.11	23.57	72.32	48.70	82.43	89.40
8	8.32	1.67	34.61	27.14	73.42	60.65	50.76	50.75
9	12.55	12.39	48.56	9.81	63.95	24.95	8.05	21.35
10	5.45	13.78	39.33	1.24	67.32	22.65	13.95	42.90
11	−4.86	−3.83	0.72	52.67	68.00	57.70	11.14	95.35
12	1.59	1.28	26.67	25.38	56.05	36.55	50.10	20.55
13	2.95	13.67	61.61	47.67	87.68	72.15	62.38	94.40
14	−2.95	11.56	22.61	22.95	33.42	53.00	89.29	15.85
15	2.73	3.00	43.28	32.76	58.32	72.55	72.43	26.90
16	−3.77	−0.94	64.44	49.19	71.79	81.00	38.76	5.85
17	−4.41	−12.72	66.44	35.19	63.42	73.50	21.86	72.60
18	−3.00	−6.56	40.00	11.38	61.63	27.30	21.62	13.90
19	1.05	1.17	33.06	17.81	73.05	60.25	30.52	50.70
20	0.45	−0.67	1.94	33.67	70.79	63.65	19.57	28.45
21	7.50	5.33	42.67	37.95	71.21	62.35	19.00	15.50
22	−6.86	0.17	35.94	40.33	78.63	91.55	18.52	62.20

Note. Items are listed in descending order of chunking scores from Experiment 1 (that is, the same order as Tables 1 and 2).

and sufficiency of the intermediate links (see Table 4 for item means for all measures). The dependent measure for this regression was a composite transitivity measure, formed by converting the item means for the $A \rightarrow C$ causal ratings (from Experiment 2) and $A \rightarrow C$ false alarm rates (from Experiment 3) to z-scores, and averaging these scores. These two measures of transitivity were highly correlated, $r(20) = .77$, $p < .001$, suggesting that these measures tapped into the same underlying construct.

We tested the role of the explanatory variables by conducting a series of regressions, adding the explanatory variables stepwise (see Table 5). In the first regression, chunking was strongly predictive of transitivity, $b = 0.38$, $SE = 0.05$, $p < .001$. In the second regression, we also included the $A \rightarrow B$ and $B \rightarrow C$ causal ratings from Experiment 2. The relationship between chunking and transitivity held up after these adjustments, $b = 0.41$, $SE = 0.06$, $p < .001$, and the causal strengths of the intermediate links did not predict transitivity (ps > .10). In the third regression, we also entered the Explanation ratings from Experiment 1 and six variables measured in Experiment 4—ΔP_{AB}, ΔP_{BC}, $\Delta P_{AC|B}$, $\Delta P_{AC|\sim B}$, Necessity(AB), and Necessity(BC). We did not enter the sufficiency ratings in this step because including these variables created a multicollinearity problem (see below for details); excluding the sufficiency ratings from the model resulted in acceptable multicollinearity diagnostics (tolerance was greater than .25 for each predictor). Once again, the relationship between chunking and transitivity held up after adjusting for all these variables, $b = 0.41$, $SE = 0.10$, $p = .002$. This shows that petering out, alternative causal pathways, and lack of necessity cannot explain why some chains are less transitive, and chunking still strongly predicts transitivity after adjusting for these variables.

The sufficiency ratings created a multicollinearity problem in part because the $B \rightarrow C$ sufficiency ratings were negatively correlated with chunking, $r(20) = -.40$, $p = .062$. Neither $A \rightarrow B$ nor $B \rightarrow C$ sufficiency was positively associated with transitivity—this correlation was nonsignificant for $A \rightarrow B$ sufficiency, $r(20) = -.15$, $p = .52$, and signifi-

Table 5
Regression analysis predicting transitivity scores

Predictor	Step One	Step Two	Step Three
Chunking (Experiment 1)	0.38 (0.05)***	0.41 (0.06)***	0.41 (0.10)**
AB causal ratings (Experiment 2)		0.26 (0.19)	0.20 (0.30)
BC causal ratings (Experiment 2)		−0.33 (0.38)	0.48 (0.71)
B explains *AC* (Experiment 1)			−0.12 (0.35)
$\Delta P_{AC\|B}$ (Experiment 4A)			−0.01 (0.02)
$\Delta P_{AC\|\sim B}$ (Experiment 4A)			0.00 (0.03)
ΔP_{AC} (Experiment 4B)			0.00 (0.01)
ΔP_{BC} (Experiment 4B)			−0.01 (0.01)
Necessity(*AB*) (Experiment 4D)			0.00 (0.01)
Necessity(*BC*) (Experiment 4D)			−0.01 (0.01)

Note. Table entries are the unstandardized coefficients (and *SE*s) in a linear regression predicting the composite transitivity measure (the mean of z-transformed $A \rightarrow C$ causal ratings and $A \rightarrow C$ false memory from Experiments 2 and 3), calculated for each item.
*$p < .05$, **$p < .01$, ***$p < .001$.

cantly negative for $B \rightarrow C$ sufficiency, $r(20) = -.54$, $p = .009$. A lack of sufficiency therefore cannot account for our intransitivity, because the sufficiency of the intermediate links was, if anything, *negatively* correlated with our transitivity scale.

An additional issue concerns our participants' violations of the causal Markov condition. These violations were measured by $\Delta P_{AC|B}$ and $\Delta P_{AC|\sim B}$, and were not significantly associated with transitivity in our regression analyses. Although it is specifically negative values of $\Delta P_{AC|B}$ and $\Delta P_{AC|\sim B}$ that would normatively lead to intransitive judgments (i.e., negative values of ΔP_{AC}), another possibility is that participants made intransitive judgments when the Markov condition was violated, regardless of whether the violation was positive or negative. In that case, one would expect the *absolute values* of $\Delta P_{AC|B}$ and $\Delta P_{AC|\sim B}$ to predict transitivity. To test this possibility, we conducted a regression analysis, parallel to step three in Table 5 but replacing $\Delta P_{AC|B}$ and $\Delta P_{AC|\sim B}$ with their absolute values. The results of this regression are nearly identical to those reported in Table 5: The absolute values of the Markov violations did not predict transitivity scores (ps > .60), but the chunking scores did ($p = .001$).

Taken together, these analyses show that the potential alternative explanations for intransitivity (petering out, alternative causal pathways, and lack of sufficiency or necessity) cannot jointly explain our chains' intransitivity: Even after adjusting for these potential explanatory variables, the chunking of causal mechanisms was strongly associated with transitivity.

6. Experiment 5

In our final experiment, we aimed to replicate the effect of schematization on transitivity using a new set of items. Specifically, we tested the possibility that schematization (or lack thereof) of a causal chain can determine transitivity so strongly that it can produce counterexamples to the most empirically well-tested account of causal transitivity—the saliency condition (Bonnefon et al., 2008, 2012). Recall that the saliency condition holds for a causal chain $A \rightarrow B \rightarrow C$ if and only if A is necessary for B. This condition could influence transitivity because A's necessity for B means that B implies A. Since B implies A, $B \rightarrow C$ would imply $A \rightarrow C$. Thus, it would be especially powerful evidence for the influence of schemas on transitivity if schematized causal chains where A is not necessary for B (e.g., item 9 in Table 1, where Tess ate rancid pork, causing her to develop a minor fever, causing her to go to the doctor) are judged more transitive than unschematized chains where A is necessary for B (e.g., item 14, where the wind blew, causing Ivan's hat to blow away, causing Ivan to buy a new hat). We refer to the first kind of counterexample as schematized/nonsalient chains, and the second kind as unschematized/salient chains. While a few items used in Experiments 1–4 were counterexamples of one of these types, it is unclear how widespread such counterexamples are. To show that schematization can lead to many of both kinds of counterexamples, we developed seven items of each type for Experiment 5, which are summarized in Table 6.

Table 6
Stimuli and results from Experiment 5

Item	Event A	Event B	Event C	Experiment 5A				Experiment 5B	
				$A \rightarrow B$	$B \rightarrow C$	$A \rightarrow C$	$A \rightarrow B$ Necessity	Explanation	Chunking
Schematized/nonsalient chains									
1	A child sprayed water at the dog, then	The dog got wet, then	The dog shook the water off its fur.	8.82	8.68	7.93	17.86	7.62	7.64
2	Someone hit Dane's car with a bat, then	Dane's car body got damaged, then	Dane took his car to the shop.	8.61	8.61	8.14	7.89	7.45	7.48
3	Selma touched a hot branding iron, then	Selma burned her hand, then	Selma swore.	8.93	8.21	8.00	12.21	7.55	6.88
4	Alex hit his hand with a hammer, then	Alex's hand ached, then	Alex put pain relief cream on his hand.	8.89	8.71	7.82	8.00	7.28	6.75
5	Betty's license plate was upside-down, then	Betty's car was noticed by the police, then	Betty got pulled over.	8.36	8.18	8.36	4.21	7.83	6.54
6	John misread the subway map, then	John got lost, then	John had to ask for directions.	8.68	8.64	8.14	18.50	7.48	6.38
7	There was a deer in the road, then	Arthur swerved his car, then	Arthur got into a car accident.	8.61	8.43	7.36	19.36	7.79	4.65
Unschematized/salient chains									
8	Bruce was tired, then	Bruce yawned, then	Bruce covered his mouth.	8.54	8.36	4.89	79.00	7.90	3.15
9	Paul thought something was really funny, then	Paul laughed very hard, then	Paul was unable to breathe.	8.75	8.14	5.68	88.18	7.90	2.63
10	Matt was hungry, then	Matt drove to a restaurant downtown, then	Matt had to find parking.	8.18	8.14	4.89	81.64	8.38	1.93

(continued)

Table 6. (continued)

Item	Event A	Event B	Event C	Experiment 5A				Experiment 5B	
				$A \rightarrow B$	$B \rightarrow C$	$A \rightarrow C$	$A \rightarrow B$ Necessity	Explanation	Chunking
11	Eric's hair was getting long, then	Eric went to the barber shop, then	Eric sat in a chair.	8.61	7.21	3.64	81.07	8.28	1.86
12	Greg lost a lot of weight, then	Greg decided to throw out his plus-size clothes, then	Greg had more trash.	8.36	8.43	5.07	78.64	7.48	1.60
13	It was hot outside, then	Theresa turned the air on high in the car, then	Theresa was unable to hear the radio.	8.64	8.46	3.89	92.07	7.66	1.48
14	Becky craved a cigarette, then	Becky started to chew a piece of nicotine gum, then	Becky moved her jaw up and down.	8.75	8.43	5.07	87.14	8.59	1.38

In Experiment 5A, we tested the saliency account for these chains. We collected causal ratings, to verify that $A \rightarrow C$ would be judged more causal for the schematized/nonsalient chains than for the unschematized/salient chains, and necessity ratings, to verify that $A \rightarrow B$ necessity would be judged higher for the unschematized/salient chains than for the schematized/nonsalient chains. Obtaining both of these differences would dissociate $A \rightarrow B$ necessity (or saliency) from $A \rightarrow C$ causal ratings. In addition, we collected $A \rightarrow B$ and $B \rightarrow C$ causal ratings as in Experiment 2, to ensure that differences in the strength of the intermediate links could not be responsible for the dissociation. To test whether the $A \rightarrow C$ ratings in Experiment 5A can be explained by schematization, Experiment 5B adopted the measures from Experiment 1. That is, we asked to what extent it was necessary to explicitly mention *B* in explaining to another person how *A* led to *C*, anticipating that these ratings would be higher for the unschematized items because the causal chain is not chunked into one unit. We also asked participants to judge the extent to which *B* explained why *A* led to *C* for each chain, to test whether the network-based definition of mechanism accounts for transitivity in this new set of chains.

6.1. Methods

We recruited 60 participants for Experiment 5 (n = 30 each for Experiments 5A and 5B). Two participants from Experiment 5A and one participant from Experiment 5B were excluded because they provided random responses on noncausal filler chains.

For Experiment 5A, each participant completed a causal ratings task and a necessity judgment task, in a counterbalanced order. The causal ratings task was similar to Experiment 2, except that participants were asked about the 14 test chains listed in Table 6 (seven schematized/nonsalient and seven unschematized/salient chains) and seven of the noncausal filler chains used in Experiments 1 and 2. That is, participants saw each of the 21 chains, in the form "*A* occurred, then *B* occurred, then *C* occurred" and rated "To what extent would you say that: [*X*] causes [*Y*]?" where *X* and *Y* were filled in with *A* and *B*, *B* and *C*, and *A* and *C*, in that order. Ratings were completed on a 9-point scale (1: "definitely would not"; 5: "unsure"; 9: "definitely would"). Causal ratings for all three links were elicited on a single screen, and each chain was presented on a separate screen in a random order.

In the necessity judgment task, participants answered, "Consider 100 cases in which *B* occurs. In how many cases was this caused by *A*?" for each of the 14 test chains on a sliding scale from 0 to 100. These questions were asked only of the $A \rightarrow B$ link because the saliency condition applies only to this link, and they were presented on separate screens in a random order.

Experiment 5B was similar to Experiment 1, except that participants were asked about the test and filler chains used in Experiment 5A. For each chain, participants were first asked to rate the extent to which *B* explained why *A* led to *C*. If the participant responded above the midpoint (5) to the *Explains* question, then he or she was asked to rate the extent to which *B* needed to be explicitly mentioned in explaining how *A* led to *C*. These

scores were reverse-coded to create the *Chunking* measure. The procedure and wordings of the dependent measures were identical to Experiment 1.

6.2. Results and discussion

Our two main hypotheses were that transitivity would be predicted by chunking (as in our previous experiments) and that transitivity would be dissociated from $A \rightarrow B$ necessity for this set of items. Both predictions were supported by linear regression analyses on the mean ratings for each of the 14 test items (see Table 6 for item means). Because Chunking and $A \rightarrow B$ were strongly correlated by design, $r(12) = -.95$, $p < .001$, separate regressions were used to avoid multicollinearity in testing these predictions.

To test whether chunking predicted transitivity, we conducted a linear regression with $A \rightarrow C$ causal ratings as the dependent variable, and $A \rightarrow B$ and $B \rightarrow C$ causal ratings (from Experiment 5A) and the explanation and chunking ratings (from Experiment 5B) as predictors. As in previous experiments, chunking was strongly predictive of transitivity (as measured by $A \rightarrow C$ causal ratings), $b = 0.67$, $SE = .09$, $p < .001$. None of the other predictor variables significantly predicted $A \rightarrow C$ causal ratings (ps $> .10$). This replicates our previous finding that chunking was associated with transitivity, even after accounting for the strength of the intermediate links. Once again, this result is consistent with schema-based representations, but it is difficult to explain with network representations.

To test whether $A \rightarrow B$ necessity could be dissociated from transitivity, we conducted a linear regression with $A \rightarrow C$ causal ratings as the dependent variable, and $A \rightarrow B$ causal ratings, $B \rightarrow C$ causal ratings, explanation ratings, and $A \rightarrow B$ necessity as predictors. As anticipated, $A \rightarrow B$ necessity was negatively associated with $A \rightarrow C$ causal strength for this set of items, $b = -0.04$, $SE = 0.01$, $p < .001$, but explanation ratings and $A \rightarrow B$ and $B \rightarrow C$ causal ratings did not significantly predict $A \rightarrow C$ causal ratings (ps $> .08$). This shows that the saliency condition—which is satisfied when A is necessary for B—can be dissociated from transitivity: Causal chains can be transitive without satisfying the saliency condition (if they are relatively schematized), and causal chains can satisfy the saliency condition without being transitive (if they are relatively unschematized).

For this set of items, saliency was a negative predictor of transitivity, whereas schematization was a positive predictor—consistent with schema-based representations. Note that this evidence only counts against a strong version of the saliency account on which saliency is necessary or sufficient (or both) for intransitivity, as we chose these items with the goal of dissociating these variables. Indeed, all else being equal, chains violating the saliency condition are less transitive than chains satisfying this condition (Bonnefon et al., 2008, 2012). The current results highlight boundary conditions on the saliency condition, however, and show that schematization can be the more influential factor when saliency and schematization are in conflict. Although these results cannot establish how prevalent such cases are, they do show that the saliency condition is neither necessary nor sufficient for transitivity, and add to the previous results in reaffirming the relationship between schematization and transitivity with a new set of items.

7. General discussion

The present experiments contribute in two ways to our understanding of how causal knowledge is represented and used. The primary issue we examined here is whether causal relations are represented as continuous networks of causal influence or as isolated causal chunks or islands. In examining this issue, we also showed that people sometimes make intransitive causal judgments, endorsing "*A* causes *B*" and "*B* causes *C*" while refusing to endorse "*A* causes *C*." In the following two sections, we briefly summarize these main contributions.

7.1. Causal networks or causal islands?

Knowledge of causal mechanisms is essential for making sense of the events around us —for determining their causes, for predicting their effects, and for planning interventions. Mechanism information is especially important for pruning the space of candidate causes, so that we can infer causal structure in the face of an infinite hypothesis space (Ahn & Kalish, 2000). However, previous empirical and theoretical work has not specified how mechanisms are mentally represented. In this article, we considered two possible organizations of causal knowledge—a *network representation*, on which causes are connected to their effects via webs of influence akin to Bayesian networks (e.g., Gopnik et al., 2004), and a *schema representation*, on which particular mechanisms are discretely stored in informational islands and mechanisms sharing a common event would not necessarily be linked to one another.

In Experiment 1, we distinguished between two ways in which a sequence of three events can fail to constitute a causal mechanism—one consistent with a network representation, and one inconsistent. First, a sequence could fail to be conceptualized as a causal mechanism simply because the intermediate event (*B*) is not seen as mediating or explaining the relationship between *A* and *C*. Although a sequence could fail to constitute a mechanism in this sense on the network theory (i.e., if *A* was not seen as causing *B* or *B* was not seen as causing *C*), our participants acknowledged that all our items satisfied this definition of causal mechanism. Nonetheless, our chains varied in the extent to which they constituted mechanisms in a second sense—that *A*, *B*, and *C* were stored as one chunk in semantic memory. Participants thought that for some of our chains, it would be necessary to explicitly mention *B* in explaining how *A* led to *C*, whereas for other chains, it was more obvious how *A* led to *C*, without mentioning *B*. We used this measurement as a proxy for the extent to which the chains were chunked into a single, coherent mechanism or were instead stored as two disparate mechanisms (one for the relationship between *A* and *B*, and another for the relationship between *B* and *C*). Chains like "Allison exercised for 20 min, then Allison became thirsty, then Allison drank a whole bottle of water" tended to be stored in one chunk, and thus participants judged that one does not need to mention that Allison became thirsty. In contrast, chains like "Francine had sex, then Francine became pregnant, then Francine experienced nausea" were more likely to

be stored in two chunks, and participants judged that one needs to mention that Francine became pregnant to explain how the initial event led to the final event. Given that these sequences are both seen as causal chains (with *A* causing *B* and *B* causing *C*), a network representation does not straightforwardly accommodate discretization of this sort.

Although these findings alone demonstrate that there is variation in the extent to which causal chains are schematized, it would be more convincing evidence that causal mechanisms are represented discretely if we could also show that this representation has consistent downstream consequences for causal inference. We used the phenomenon of *causal transitivity* to provide such evidence. On a network representation, the causal strength of a causal chain should be a function of the number and strength of the links (Baetu & Baker, 2009; see also Anderson, 1983 on traversing links in other kinds of semantic networks). Therefore, for two chains $A \rightarrow B \rightarrow C$ and $X \rightarrow Y \rightarrow Z$ where $A \rightarrow B$ and $X \rightarrow Y$ are equally strong and where $B \rightarrow C$ and $Y \rightarrow Z$ are equally strong, the transitive $A \rightarrow C$ and $X \rightarrow Z$ links should be equally strong. However, Experiments 2 and 3 showed that this is not the case. To the extent that a chain was found to be chunked in Experiment 1, people were more willing to infer that *A* causes *C*, holding constant the strength of the intermediate links. This effect held up both in direct causal ratings (Experiment 2) and in a recognition memory task (Experiment 3). For example, when Allison exercised, became thirsty, and drank water, participants tended to schematize this as one chunk in Experiment 1, and also tended to make the transitive inference (Allison exercising caused her to drink water) in Experiments 2 and 3. In contrast, when Francine had sex, became pregnant, and experienced nausea, participants tended to divide this chain into two chunks in Experiment 1 and were much less likely to make the transitive inference (Francine having sex caused her to experience nausea) in Experiments 2 and 3.

We studied inferences principally at the token level (e.g., Francine's sex causing her to become nauseous) rather than at the type level (e.g., sex in general causing nausea), but the predictions of the network and schema theories also apply at the type level. However, because people rely more on statistical evidence for evaluating type-causal relationships (Johnson & Keil, *in prep*), a statistics-driven Bayesian network approach could potentially account more accurately for transitivity judgments. Furthermore, some of the alternative accounts rely on type-level factors (e.g., necessity and sufficiency of the intermediate links), so it is possible that once these factors are accounted for, no intransitivity would remain at the type level. As a test of this possibility, we collected conditional probability judgments of $P(C|A)$ and $P(C|\sim A)$ at the type level using the same procedure as Experiment 4B, to calculate $\Delta P_{AC} = P(C|A) - P(C|\sim A)$. A regression parallel to step three in Table 5 found that schematization also predicted these ΔP scores after adjusting for the other explanatory variables, $b = 4.40$, $SE = 1.27$, $p = .002$, consistent with the schema-based approach but once again in tension with the network-based approach.

7.2. Are causal relations transitive?

The issue of causal transitivity is also important in its own right, as a way to combine premises in causal reasoning. Although transitive inferences have been found using more

artificial stimuli (Ahn & Dennis, 2000; Goldvarg & Johnson-Laird, 2001; Von Sydow et al., 2009), less work has examined inference patterns for chains of familiar events. It is unlikely a priori that all causal chains should be judged transitive, because several normative reasons for intransitivity have been documented (Hitchcock, 2001), which we termed *threshold effects*, *incompatible aspects*, *petering out*, *alternative causal pathways*, and *lack of necessity or sufficiency*. To see whether these previously documented reasons can account for all causal intransitivity, we tested these accounts for our set of causal chains.

In Experiment 2, we addressed the possibility of *threshold effects* or *incompatible aspects* as explanations for our findings (Hausman, 1992; McDermott, 1995; Paul, 2000; Schaffer, 2005). First, a *threshold effect* occurs when *A* affects the value of *B* and *B* affects the value of *C*, but *A* does not affect the value of *B* sufficiently to affect the value of *C*. However, in Experiment 2, the causal chain was described at the token level, so the value of *B* was the same when *A* caused *B* and when *B* caused *C* within each chain. Threshold effects are thus not possible explanations for the intransitive chains. Second, a causal chain can be intransitive due to *incompatible aspects* when the property of *B* modified by *A* is not the same property of *B* relevant for modifying *C*. However, this also seems unlikely to explain our findings. Most of our chains only indicated one property of *B* which could be modified (i.e., whether *B* occurs or not), so the aspect of *B* modified by *A* must be the same as the aspect of *B* modifying *C*. For example, having sex causes one to be pregnant (rather than not pregnant), and being pregnant (rather than not pregnant) causes one to experience nausea. The aspect of the intermediate event under consideration appears to be the same in both links of each causal chain, so incompatible aspects appear unable to explain our results.

In Experiment 4, we measured the contributions of *alternative causal pathways* (Eells & Sober, 1983), *petering out* (Lowe, 1980), and *lack of necessity or sufficiency* (e.g., Bonnefon et al., 2008) to our intransitive chains. An *alternative causal pathway* occurs when *A*'s occurrence activates multiple pathways—one causing *C* and one preventing *C*—leading to a weak contingency between *A* and *C*. *Petering out* occurs when the probabilistic strengths of the intermediate links are relatively weak, so the overall contingency between *A* and *C* is so weak as to be considered noncausal. A *lack of necessity or sufficiency* could potentially account for the differences in transitivity if the intermediate links of some chains were more necessary or sufficient than others. In Experiment 4, we measured the probabilistic strength of each intermediate link, the probabilistic strength of alternative causal pathways, and the sufficiency and necessity of each link. After adjusting for these variables in a multiple regression, the chunking measure from Experiment 1 continued to predict transitivity, yet none of these probability variables were significant predictors. In addition, Experiment 5 examined a specific proposal by Bonnefon et al. (2008, 2012), emphasizing the role that $A \rightarrow B$ necessity plays in judgments of transitivity. Using 14 new chains, we found that $A \rightarrow B$ necessity judgments can dissociate from transitivity, demonstrating that $A \rightarrow B$ necessity is neither necessary nor sufficient for transitivity, but that schematization nonetheless predicts transitivity. It is certainly possible, however, that the necessity of the $A \rightarrow B$ link plays a subtler role in influencing transitive inferences.

None of this is to deny that these normative factors can lead to intransitive causal judgment. Indeed, incompatible aspects (Hagmayer et al., 2011) and lack of $A \rightarrow B$ necessity (Bonnefon et al., 2012) have been empirically shown to result in intransitive causal judgments. Rather, these factors cannot account for the intransitivity found in the current experiments, but lack of schematization can.

7.3. Why are some causal chains schematized?

One issue that is unresolved by these experiments is what causes some links to be schematized together but others not to be. It is not simply a matter of some chains having stronger causal links—this possibility was ruled out in Experiments 2 and 4. Instead, other factors must be driving schematization. Here, we consider three factors that could potentially contribute—the homogeneity of mechanisms, temporal contiguity, and temporal discreteness.

One possible factor is the homogeneity of the mechanisms underlying the $A \rightarrow B$ link and the $B \rightarrow C$ link. For example, in the relatively schematized chain 5 ("Melissa was outside in warm weather, then her body temperature rose, then her clothes were soaked with sweat"), it seems that a single physiological mechanism—the body's homeostasis process—underlies both links in the chain. But in the case of sex, pregnancy, and nausea, one could argue that the physiological mechanisms connecting sex and pregnancy are very different from the physiological mechanisms connecting pregnancy and nausea, and it is merely coincidental that these two physiological mechanisms are connected in the middle by the same event. However, this explanation potentially risks circularity without having an independent definition of homogeneity, as our intuitions about the homogeneity of these mechanisms could instead be driven by schematization rather than the reverse. Indeed, this problem is similar to the classic and still-unresolved problem of what makes concepts coherent (or homogeneous). For instance, we have concepts such as emeralds, but we do not have concepts such as *emerubies*—an emerald before 1997 or a ruby after 1997 (Goodman, 1955).

A second factor that could influence schematization is temporal contiguity. To the extent that two events co-occur closer together in time, they are likelier to participate in a genuinely causal relationship (e.g., Johnson & Keil, 2014; Lagnado & Sloman, 2006) and likelier to be schematized. However, it does not appear that temporal contiguity can explain much of the variability in our participants' judgments. Although some of our relatively unschematized chains (e.g., Francine having sex, becoming pregnant, and experiencing nausea) do lack temporal contiguity, other unschematized chains are highly contiguous (e.g., Karen stepping on a dog, the dog growling, and a child being scared) and some schematized chains are relatively *dis*contiguous (e.g., Pam failing to floss her teeth, having plaque on her teeth, and developing cavities). Another problem with the temporal contiguity explanation is that if A and B (or B and C) are discontiguous, making the chain $A \rightarrow B \rightarrow C$ unschematized, then the causal strength of the $A \rightarrow B$ link (or the $B \rightarrow C$ link) would have been weak as well. Yet Experiments 2 and 5 found that our participants' intransitive judgments could not be explained by the strength of these intermediate links.

A third possibility concerns the *nature* of the temporal relationships among events in the causal chains. For example, sex causing a person to become pregnant may call to mind a temporally discrete event (i.e., conception), whereas it is the temporally extended event of pregnancy that is associated with nausea. Thus, even though both intermediate causal links are very strong, the overall relationship might be less likely to be schematized due to the different time scales on which the links occur. Although this sort of explanation does not appear to apply to the majority of the unschematized chains used in this article, it may contribute to some of them and is an interesting object of future study.

We suspect that schematization is driven by a number of factors, including mechanistic homogeneity and temporal factors, as well as the frequency with which we encounter these chains in everyday experience and similarity of the links along dimensions other than causality. If schematization is indeed multiply determined, this renders the problem highly challenging, but all the more interesting for future research.

7.4. Pluralistic strategies for causal inference

Recently, many researchers across cognitive science have argued that our causal concepts and inference strategies are *pluralistic*—that is, people have a multiplicity of causal concepts and inference strategies that are deployed flexibly in context-dependent ways (e.g., Danks, 2005; Hitchcock, 2003; Lombrozo, 2010; Woodward, 2011). As we noted in the introduction, the link between schematization and transitivity follows from a commonly used *narrative strategy* for assessing causality, wherein a reasoner assesses a causal relationship between X and Y by trying to think of a plausible story for how X would lead to Y based on their background knowledge (Kahneman & Tversky, 1982; Taleb, 2007). If people use discrete schema-based representations, we argued, they would be more likely to make a transitive inference when X and Z are stored in the same schema because this would make the narrative easier to construct. In contrast, a network representation would have no resources to explain differences in transitivity unless the intermediate links differed in causal strength, since the network representation is not discrete.

However, this explanation implies that schematization and transitivity would be closely related only when people are using a narrative strategy. Indeed, this observation is consistent with previous research, where transitive inferences have been found in cases where a narrative strategy was unavailable, but other strategies could be used instead. In studies using artificial stimuli where a narrative was missing but covariation information was available, enabling use of a *statistical strategy*, people robustly inferred high $A \rightarrow C$ covariation from high $A \rightarrow B$ and $B \rightarrow C$ covariation (e.g., Ahn & Dennis, 2000; Baetu & Baker, 2009), even to the point of inferring illusory correlations where in fact no correlation exists (Von Sydow et al., 2009). And in studies using randomly constructed, abstract premises (e.g., "Obedience causes motivation to increase" and "Increased motivation causes eccentricity"), people used a *rule-based strategy* to combine the premises, drawing transitive conclusions ("Obedience causes eccentricity"; Goldvarg & Johnson-Laird, 2001).

In those studies using artificial stimuli, participants' use of these strategies were successfully isolated from their background knowledge, demonstrating that both a statistical strategy and a rule-based strategy can result in transitive inferences. Yet, because we generally rely on prior knowledge for making causal inferences when it is available, it is likely that we more often adopt a narrative strategy in everyday causal inference (Kahneman & Tversky, 1982; Taleb, 2007; see also Ahn & Kalish, 2000). Consequently, intransitive judgment may be relatively common in everyday causal thinking, in cases where disparate schemas collide.

8. Conclusion

Causal knowledge is a primary tool we use to make sense of the flow of experience. Despite this sense of flow, however, we have argued here that we break up these causal relations into discrete units, into schematized causal mechanisms. This organization of causal knowledge may be helpful in clustering stable causal patterns together, in much the same way that categories are useful ways of organizing reliably co-occurring features. Such ways of breaking the world into smaller pieces may help us to make sense of experience in a way that is manageable, given our cognitive limits.

Acknowledgments

We thank Angie Johnston, Frank Keil, and Joshua Knobe for their comments on earlier versions of this manuscript; Christian Luhmann and Jessecae Marsh for helpful discussion in the initial stages of this research; Phillip Wolff for suggesting the method used in Experiment 3; the members of the Thinking Lab for their suggestions; and audiences at Yale University and Northwestern University for helpful discussion. Support for this research was provided by grant R01 MH057737 and R01 HG007653 from the National Institutes of Health, awarded to the second author.

Note

1. We supplemented this item-level analysis with an additional analysis at the level of individual participants. For each participant, we calculated the partial correlation between that participant's ratings of $A \rightarrow C$ for each chain and each chain's mean chunking score from Experiment 1, adjusting for that participant's $A \rightarrow B$ and $B \rightarrow C$ causal judgments. Three participants could not be included in this analysis because they rated all the $A \rightarrow B$ or $B \rightarrow C$ links at ceiling. Among the remaining participants, their Fisher-transformed partial correlations were significantly greater than 0, $t(26) = 7.99$, $p < .001$, with a mean inverse-transformed partial correlation of $r = .44$. Thus, the chunking ratings made by a separate group of participants in

Experiment 1 predicted the $A \rightarrow C$ ratings of individual participants in Experiment 2. Although we do not report subject-level analyses for subsequent experiments, subject-level analyses or raw data are available upon request.

References

Ahn, W., & Bailenson, J. (1996). Causal attribution as a search for underlying mechanisms: An explanation of the conjunction fallacy and the discounting principle. *Cognitive Psychology*, *31*, 82–123.

Ahn, W., & Dennis, M. J. (2000). Induction of causal chains. In L. R. Gleitman, & A. K. Joshi (Eds.), *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp. 19–24). Mahwah, NJ: Erlbaum.

Ahn, W., & Kalish, C. W. (2000). The role of mechanism beliefs in causal reasoning. In F. C. Keil & R. A. Wilson (Eds.), *Explanation and cognition* (pp. 199–226). Cambridge, MA: MIT Press.

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.

Alba, J. W., & Hasher, L. (1983). Is memory schematic? *Psychological Bulletin*, *93*, 203–231.

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, *22*, 261–295.

Baetu, I., & Baker, A. G. (2009). Human judgments of positive and negative causal chains. *Journal of Experimental Psychology: Animal Behavior Processes*, *35*, 153–168.

Bartlett, F. C. (1932). *Remembering: An experimental and social study*. Cambridge, UK: Cambridge University Press.

Baumrind, D. (1983). Specious causal attributions in the social sciences: The reformulated stepping-stone theory of heroin use as exemplar. *Journal of Personality and Social Psychology*, *45*, 1289–1298.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*, 289–300.

Björnsson, G. (2006). How effects depend on their causes, why causal transitivity fails, and why we care about causation. *Philosophical Studies*, *133*, 349–390.

Bonnefon, J.-F., Da Silva Neves, R., Dubois, D., & Prade, H. (2008). Predicting causality ascriptions from background knowledge: Model and experimental validation. *International Journal of Approximate Reasoning*, *48*, 752–765.

Bonnefon, J.-F., Da Silva Neves, R., Dubois, D., & Prade, H. (2012). Qualitative and quantitative conditions for the transitivity of perceived causation: Theoretical and experimental results. *Annals of Mathematics and Artificial Intelligence*, *64*, 311–333.

Bransford, J. D., Barclay, J. R., & Franks, J. J. (1972). Sentence memory: A constructive versus interpretive approach. *Cognitive Psychology*, *3*, 193–209.

Brewer, W. F. (1977). Memory for the pragmatic implications of sentences. *Memory & Cognition*, *5*, 673–678.

Broadbent, A. (2012). Causes of causes. *Philosophical Studies*, *158*, 457–476.

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

Bullock, M., Gelman, R., & Baillargeon, R. (1982). The development of causal reasoning. In W. J. Friedman (Ed.), *The developmental psychology of time* (pp. 209–254). New York: Academic Press.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*, 55–81.

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

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*, 407–428.

Danks, D. (2005). The supposed competition between theories of human causal inference. *Philosophical Psychology*, *18*, 259–272.

Dietrich, E., & Markman, A. B. (2003). Discrete thoughts: Why cognition must use discrete representations. *Mind & Language*, *18*, 95–119.

Eells, E., & Sober, E. (1983). Probabilistic causality and the question of transitivity. *Philosophy of Science*, *50*, 35–57.

Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science*, *171*, 303–306.

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

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*, 684–704.

Glymour, C., & Cheng, P. W. (1998). Causal mechanism and probability: A normative approach. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 295–313). Oxford, UK: Oxford University Press.

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.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review*, *111*, 3–32.

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

Hagmayer, Y., Meder, B., von Sydow, M., & Waldmann, M. R. (2011). Category transfer in sequential causal learning: The unbroken mechanism hypothesis. *Cognitive Science*, *35*, 842–873.

Hausman, D. M. (1992). Thresholds, transitivity, overdetermination, and events. *Analysis*, *52*, 159–163.

Hitchcock, C. (2001). The intransitivity of causation revealed in equations and graphs. *The Journal of Philosophy*, *98*, 273–299.

Hitchcock, C. (2003). Of Humean bondage. *The British Journal for the Philosophy of Science*, *54*, 1–25.

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs: General and Applied*, *79*, 1–17.

Johnson, S. G. B., & Keil, F. C. (2014). Causal inference and the hierarchical structure of experience. *Journal of Experimental Psychology: General*, *143*, 2223–2241.

Kahneman, D., & Tversky, A. (1982). The simulation heuristic. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 201–208). Cambridge, UK: Cambridge University Press.

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

Lagnado, D. A., Waldmann, M. R., Hagmayer, Y., & Sloman, S. A. (2007). Beyond covariation: Cues to causal structure. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 154–172). Oxford, UK: Oxford University Press.

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

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

Livingston, K. R., Andrews, J. K., & Harnad, S. (1998). Categorical perception effects induced by category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 732–753.

Lombrozo, T. (2010). Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, *61*, 303–332.

Lowe, E. J. (1980). For want of a nail. *Analysis*, *40*, 50–52.

Mackie, J. L. (1965). Causes and conditions. *American Philosophical Quarterly*, *2*, 245–264.

Markman, A. B. (1999). *Knowledge representation*. Mahwah, NJ: Erlbaum.

McDermott, M. (1995). Redundant causation. *The British Journal for the Philosophy of Science*, *46*, 523–544.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81–97.

Murphy, G. L. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Park, J., & Sloman, S. A. (2013). Mechanistic beliefs determine adherence to the Markov property in causal reasoning. *Cognitive Psychology*, *67*, 186–216.

Paul, L. A. (2000). Aspect causation. *The Journal of Philosophy*, *97*, 235–256.

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.)

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*, 573–605.

Rosen, D. A. (1978). In defense of a probabilistic theory of causality. *Philosophy of Science*, *4*, 604–613.

Schaffer, J. (2005). Contrastive causation. *The Philosophical Review*, *114*, 327–358.

Schank, R., & Abelson, R. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. New York, NY: Psychology Press.

Shultz, T. R. (1982). Rules of causal attribution. *Monographs of the Society for Research in Child Development*, *47*, 1–51.

Sloman, S. A. (1994). When explanations compete: The role of explanatory coherence on judgements of likelihood. *Cognition*, *52*, 1–21.

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

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

Taleb, N. N. (2007). *The black swan: The impact of the highly improbable*. New York: Random House.

Von Sydow, M., Meder, B., & Hagmayer, Y. (2009). A transitivity heuristic of probabilistic causal reasoning. In N. A. Taatgen & H. van Rign (Eds.), *Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society* (pp. 803–808). Austin, TX: Cognitive Science Society.

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.

Walsh, C. R., & Sloman, S. A. (2011). The meaning of *Cause* and *Prevent*: The role of causal mechanism. *Mind & Language*, *26*, 21–52.

Woodward, J. (2011). A philosopher looks at tool use and causal understanding. In T. McCormack, C. Hoerl, & S. Butterfill (Eds.), *Tool use and causal cognition* (pp. 18–50). Oxford, UK: Oxford University Press.

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