Causal conflicts produce domino effects

Sangeet Khemlani

Naval Research Laboratory

P.N. Johnson-Laird

Princeton University

New York University

June 9th, 2020

Corresponding author:

Sangeet Khemlani

Navy Center for Applied Research in Artificial Intelligence

Naval Research Laboratory

4555 Overlook Ave SW

Washington, DC 20375

Email: sunny.khemlani@nrl.navy.mil

Abstract

Inconsistent beliefs call for revision – but which of them should individuals revise? A long-standing view is that they should make minimal changes that restore consistency. An alternative view is that their primary task is to explain how the inconsistency arose. Hence, they are likely to violate minimalism in two ways: they should infer more information than is strictly necessary to establish consistency, and they should reject more information than is strictly necessary to establish consistency. Previous studies corroborated the first effect: reasoners use causal simulations to build explanations that resolve inconsistencies. Here, we show that the second effect is true too: they use causal simulations to reject more information than is strictly necessary to establish consistency. When they abandon a cause, the effects of the cause topple like dominos: Reasoners tend to deny the occurrence of each subsequent event in the chain. Four studies corroborated this prediction.

Keywords: inconsistency, bridging inferences, domino effects, causal reasoning, minimalism, mental models

1. Introduction

When individuals run into a fact contradicting earlier information, they need to change their mind. They have to modify one or more beliefs to restore consistency. But, which beliefs should they change? The predominant view in philosophy and artificial intelligence is that they should preserve as much of the old information as they can (Gärdenfors, 1988; James, 1907; Levi, 1991; Quine, 1992). For instance, Harman (1986, p. 59) writes, "It seems that in changing one's view one should make minimal changes, both in adding new beliefs and in eliminating beliefs, for example, in order to get rid of an inconsistency in one's view". Such *minimalism* postulates that one should not change more beliefs than are necessary to accommodate the new fact. Gärdenfors offers the following illustration:

> "...assume that you believe, firstly, that all persons in the bank after four o'clock yesterday were employees and, secondly, that all employees are honest. If you then obtain compelling evidence that money was stolen in the bank during this time, you must retract one of these beliefs. *It seems irrational, however, to retract both of these beliefs, since this would involve an unnecessary loss of information*" (1982, p. 137, our italics).

In other words, minimalist accounts predict that individuals should refrain from rejecting $n + 1$ propositions when the rejection of n propositions yields a consistent set. Systems in artificial intelligence have adopted this minimalist constraint as a rational way to resolve inconsistencies (Baltag, Gierasimczuk, & Smets, 2011; Coste-Marquis, Konieczny, Mailly, & Marquis, 2014; Euzenat, 2015; Fermé & Hansson, 2011).

Recent evidence suggests that human reasoners are not minimalists. When they encounter an inconsistency, their chief task appears to be to make sense of how it arose. They aim to explain it, and their explanations often violate minimalism (Legrenzi & Johnson-Laird, 2005; Walsh & Johnson-Laird, 2009). Likewise, individuals who have an explanation in mind are faster to revise their beliefs than those who do not (Khemlani & Johnson-Laird, 2013). These results support an alternative to minimalism, i.e., the *causal simulation hypothesis*. It postulates that individuals simulate causal sequences of events to resolve inconsistencies, and that these simulations can create changes that go beyond minimality. Individuals even rate such changes as more probable than minimal ones (Johnson-Laird, Girotto, & Legrenzi, 2004).

There are two ways in which causal knowledge can make changes that transcend minimalism. The first way, as in the aforementioned studies, introduces novel properties, relations, or entities, in order to create explanations. For example, Khemlani & Johnson-Laird (2011) reported a study in which reasoners made inferences from premises, such as:

1. If a person is bitten by a viper then that person dies.

   Someone was bitten by a viper but did not die.

   What follows?

The participants spontaneously drew conclusions, such as:

2a. The person received an antidote.

 b. The person was wearing heavy clothing.

Conclusion (2a) refutes the first premise in (1) and conclusion (2b) refutes the second premise in (1). But they are not mere denials: (2a) introduces an entity (the antidote) and a relation (its reception), and (2b) likewise introduces an entity (heavy clothing) and a

relation (blocking the bite). Hence, they serve as causal explanations. The conclusions are in striking contrast to their minimal alternatives:

3a. The viper's bite was not deadly.

b. The person was not bitten by the viper.

Reasoners were less likely to make such revisions even though they restore consistency. When they had to choose one of the four options (2a, 2b, 3a, 3b), they preferred explanations to minimal changes. In sum, when individuals resolve inconsistencies, they tend to formulate explanations that violate minimalism by introducing new ideas. The new information has the side effect of eliminating or modifying the prior beliefs yielding the inconsistency (Khemlani & Johnson-Laird, 2012).

The second way to violate minimalism is to reject more information than is strictly necessary to restore consistency. In other words, reasoners who reject $n+1$ propositions when they need only reject n propositions to resolve a conflict are in violation of minimalism. No studies that we know of have tested this potential change – but causal simulation predicts it. The hypothesis postulates that individuals use their background knowledge to simulate causal relations between events if they are missing from descriptions. In the absence of background information about alternative causes for an effect, individuals tend to make a strong causal interpretation in which the cause is unique in bringing about the effect (Goldvarg & Johnson-Laird, 2001; Khemlani, Barbey, & Johnson-Laird, 2014). Hence, a fact that refutes a link in the causal chain, i.e., it is contrary to one event (or non-event) causing another, should have a domino effect in which all of the subsequent links in the chain cease to hold. In real world situations, however, reasoners may have background knowledge pertaining to possible alternative

causes for various events, and so reasoners should be less likely to interpret an assertion as refuting a causal link, and thereby creating a domino effect. Nonetheless, whenever reasoners withdraw belief in any subsequent causal link, they reject more than is necessary to restore consistency.

Consider, for instance, this scenario:

4. Sarah turned on the kitchen light.

   The bulb burst.

   Glass fell on the kitchen counter.

Individuals should make "bridging" inferences (Clark, 1975) to simulate a causal chain from turning on the light, to the bulb bursting, to its glass falling on the counter. Evidence corroborates the occurrence of such simulations, which can unfold kinematically (see, e.g., Khemlani, Barbey, & Johnson-Laird, 2014; Khemlani, Mackiewicz, Bucciarelli, & Johnson-Laird, 2013). Suppose individuals read the preceding scenario, and then learn:

5. In fact, Sarah did not turn on the kitchen light.

They are likely to cease to belief the first statement in the preceding scenario (4). That is a minimal revision. The effect of a bulb bursting can occur in the absence of the cause in the scenario, because some other cause, such as a bird flying into the bulb, can bring it about (Goldvarg & Johnson-Laird, 2001). Hence, the contradiction of a cause does not necessarily imply that its effect in a causal simulation does not occur. A minimal revision therefore does not call for its denial. (Minimalism could be modified to accommodate bridging inferences without a violation of a minimal change – but no theorist has proposed such a solution, probably because such a step renders the notion of

minimalism vacuous. We return to this point below.) The present theory postulates that when a fact refutes a link in a causal chain, the simulation of the causal chain halts at that point – it no longer propagates. And it embodies a temporal constraint: causes do not occur after their effects (Goldvarg & Johnson-Laird, 2001). So, the effects of a cause propagate from one contemporaneous event to another or forwards in time from one event to the next, but they do not propagate backwards in time. When individuals halt a causal simulation, they should tend to treat events occurring after the refuted link, rather than before it, as not occurring. They should therefore be susceptible to a domino effect in which their disbelief in an event leads to disbelief in the event that followed, which in turn, leads to disbelief in its subsequent event, and so on.

In contrast, suppose individuals read the scenario:

6. Katie switched off the washing machine.

The cat meowed.

The children came home from school.

And then they learn:

In fact, Katie did not switch off the washing machine.

They should cease to believe the contradicted event, but the scenario is a series of independent events that resist a causal simulation, and so no domino effect should occur.

In summary, minimalism does not predict any difference between causal and control scenarios of independent events. Individuals should cease to believe only the statement in the scenario that the fact refutes. Accounts of minimalism would need radical alterations to accommodate bridging inferences and domino effects in a way that did not count as violating their basic principles. In contrast, the theory of causal

simulation predicts that scenarios eliciting a simulation of a causal chain should yield a domino effect in which disbelief propagates down the causal chain from the event that the fact refutes, whereas it predicts that control scenarios, which do not elicit a simulation of a causal chain, should not yield a domino effect.

Our goal in this paper is to compare the two predictions. We carried out four experiments to test them. Experiments 1 and 2 presented participants with scenarios of three events, half of which could be simulated in a causal chain. A fact then contradicted one of the events, and their task was to judge whether each of the three statements in the scenario was true or false. Their responses revealed domino effects. Of course, contradictions should affect, not just decisions about the truth or falsity of a statement, but also degrees of belief. Recent theories postulate that estimates of subjective probabilities are more sensitive measures of beliefs than judgments of truth or falsity (e.g., Evans, 2012; Oaksford & Chater, 2013; Over, 2009). Experiment 3 accordingly called for participants to estimate the likelihood of the statements in the scenarios other than the one that the fact contradicted, and it corroborated domino effects in the estimates of likelihood. And Experiment 4 compared causal sequences and temporal sequences, and domino effects occurred only for causal sequences.

2. Experiment 1

In previous studies, participants often failed to detect inconsistencies (Otero & Kintsch, 1992), and so the present experiment used simple sentences and an obvious contradiction of one of them. On each trial, participants received a description of three separate events in a scenario. The experiment manipulated whether the description could

be construed as a causal chain of events or not. *Causal* scenarios implied that the events were in a causal sequence, though they did not use explicit causal verbs, e.g., "cause," "make," and "force". For instance, the following set of statements can elicit the simulation of a causal chain of events:

David put a book on the shelf.

The shelf collapsed.

The vase broke.

The participants then received a fourth statement that contradicted one of the three preceding statements, e.g.:

In fact, David did not put a book on the shelf.

Control scenarios presented a series of independent events that should not elicit causal simulations, e.g.:

Robert heard a creak in the hall closet.

The faucet dripped.

The lawn sprinklers started.

A fourth statement then contradicted one of the three preceding statements, e.g.,

In fact, Robert did not hear a creak in the hall closet.

Minimalism predicts that for both causal and control scenarios, individuals should cease to believe only the statement that the fact contradicts. In contrast, causal simulation predicts that for causal scenarios, reasoners should exhibit domino effects in which the initial contradiction propagates disbelief in the successive events. Experiment 1 tested these contrasting predictions. The participants' task on a given trial was to decide

whether one of the three statements in the scenario was true or false in the light of the contradiction.

2.1 Method

2.1.1 Participants. 32 participants (15 males, mean age = 31.0 years) were recruited on an online platform hosted on Amazon Mechanical Turk, and they completed the study for monetary compensation. Participation was restricted to United States residents, and repeat participation, both within and across experiments, was not allowed. None of the participants in Experiment 1, or in any subsequent experiments, had received any training in logic.

2.1.2 Design, materials, and procedure. Participants acted as their own controls and received 12 sets of 3 statements, 6 were causal scenarios and 6 were control scenarios. After the presentation of the three statements, the program revealed a fourth statement that contradicted one of the 3 preceding ones in the scenario. The contents of the problems were rotated so that the fourth statement contradicted the first statement, the second statement, and the third statement, equally often across the study as a whole. Participants then answered a question (*Did X happen?*), where *X* referred at random, but equally often, to one of the three statements. Likewise, the 12 problems appeared in a different random order for each participant.

The contents were 6 causal scenarios and 6 control scenarios (see Table A1 in the Appendix for the contents of all the experiments). As in the examples above, a named individual carries out an action. In the causal scenarios, it implies an effect, which in turn

implies a further effect. In the control scenarios, the events have no implied effects, and should not elicit a causal simulation. Each causal scenario had a matching control scenario with the same number of syllables.

The instructions to the experiment explained that it was neither an intelligence test nor a personality test, and that it concerned general patterns of thinking. The key instruction was:

> Our aim in this study is to find out how people reason about conflicts in information. On each trial, you'll be presented with some information about a series of events. Then you will be presented with factual information that conflicts with some of the events. Your task is to resolve the conflict. [The instructions gave an example of a problem.] Based on the information you're given, you will be asked to respond to the question by selecting a button for "Yes" or one for "No".

The participants were told that they were free to stop carrying out the experiment at any time.

2.2 Results and discussion

Table 1 presents the percentages of "no" responses to each of the three statements in the descriptions, i.e., the proportion of trials on which the participants judged that a statement was false. The cells in grey along the diagonals show the participants' evaluations of the statements that the facts directly contradicted. As the table shows, participants answered sensibly: they rejected directly contradicted statements more often than all the other statements (87% vs. 31%, Wilcoxon test, $z = 4.63$, $p < .0001$, Cliff's $\delta =$

.80). They also rejected statements in causal scenarios more often than those in control scenarios (63% vs. 34%, Wilcoxon test, $z = 3.49$, $p = .0005$, Cliff's $\delta = .58$). This difference reflects causal simulations. Participants rejected statements on 55% of trials in which the first statement was contradicted, on 51% of trials in which the second statement was contradicted, and on 39% of trials in which the third statement was

	Fact contradicted...		
	…statement 1	…statement 2	…statement 3
Causal scenarios			
Did event in statement 1 occur?	*87*	16	15
Did event in statement 2 occur?	**79**	*96*	23
Did event in statement 3 occur?	**70**	**73**	*88*
Control scenarios			
Did event in statement 1 occur?	*80*	13	17
Did event in statement 2 occur?	20	*81*	5
Did event in statement 3 occur?	0	27	*85*

Table 1. The percentages of "no" answers in Experiment 1 to questions about the occurrence of the events in the first, second, or third statement, depending on whether the scenario was causal or control, and on whether the fact contradicted the first, second, or third statement in the scenario. Grey italicized cells denote answers about statements that the facts contradicted, and bold cells highlight domino effects.

contradicted (Page's trend test, $z = 2.25$, $p = .025$). The results likewise revealed an interaction: participants rejected statements that occurred after a contradicted statement rather than before a contradicted statement, but they did so more often for causal scenarios than for control scenarios (Mann-Whitney test, $z = 4.73$, $p < .0001$, Cliff's $\delta = .66$). Planned comparisons elucidated the interaction: the numbers underneath the grey diagonal are larger than those above for causal scenarios (Wilcoxon test, $z = 4.20$, $p < .0001$, Cliff's $\delta = .67$), but not for control scenarios (Wilcoxon test, $z = 1.10$, $p = .27$, Cliff's $\delta = .09$). So, the domino effect occurred reliably only for causal scenarios.

The data suggested two other phenomena. First, when the first statement was contradicted, the domino effect appeared to decline, i.e., 79% of participants rejected the second statement but only 70% rejected the third statement. This effect was not reliable (Mann-Whitney test, $z = .67$, $p = .50$, Cliff's $\delta = .09$). Second, 5 out of the 32 participants accepted at least one statement that the facts directly contradicted. One explanation is that these participants were confused by having to evaluate statements that had been directly contradicted. They were told that an earlier event did not occur, and then they had to evaluate whether or not the event had occurred. They did not balk at the task, but it may have seemed odd enough to confuse them, and to make them guess. Guesswork may have infected other trials. Hence, the presence of direct contradictions may have decreased the participants' confidence in the information in the scenarios. Likewise, the experiment forced them to confront contradictions, and this experience may have increased the subjective difficulty of the task. Experiment 2 aimed to eliminate these problems: the participants evaluated only those statements that were not in direct contradiction with the facts.

3. Experiment 2

Experiment 2 used the same task and materials as the previous study. But, the participants evaluated only those events in scenarios that the facts did not contradict. For instance, given the following problem:

Harry pulled the trigger.

The gun fired.

The bullet shattered a window.

In fact, Harry did not pull the trigger.

the participants answered questions on separate trials, not about the event in the first statement, but only about the other two statements:

Did the gun fire?

Did the bullet shatter a window?

3.1 Method

3.1.1 Participants. 20 participants (10 males; mean age = 33.6) from the same population as before completed the study for monetary compensation.

3.1.2 Design, materials, and procedure. Participants evaluated 12 descriptions in 6 causal and 6 control scenarios (see Appendix). A statement of fact then contradicted one of the 3 statements in the scenario. The task, as before, was to respond to the question, *Did X happen?* where *X* referred to an event described in the first, second, or third statement in the scenario, with the additional proviso that *X* was not the statement that the fact contradicted. So, when the fact contradicted, say, the second statement in the scenario, the participants were either asked whether the event in the first statement occurred or else whether the event in the third statement occurred. The experiment counterbalanced which of the 2 non-contradicted statements to present such that participants responded to 4 problems that asked whether the first statement occurred (2 causal and 2 control), 4 problems that asked whether the second statement occurred, and 4 that asked whether the third statement occurred. The experiment presented the 12 problems in a different

random order for each participant. The materials and procedure were otherwise the same as those for Experiment 1.

3.2 Results and discussion

Table 2 presents the percentages of trials in Experiment 2 on which participants responded "no" to the question, *Did X happen?* when the fact had contradicted one of the other statements. The results replicated the corresponding results in Experiment 1. The participants rejected statements, which were not directly contradicted, on 51% of trials with causal scenarios, but on only 5% of trials with control scenarios (Wilcoxon test, $z = 3.84$, $p < .0001$, Cliff's $\delta = .87$). They rejected 44% of statements when the fact contradicted the first statement in a scenario, 30% of statements when it contradicted the second statement, and 11% of statements when it contradicted the third statement (Page's trend test, $z = 4.42$, $p < .0001$). The data again corroborated the interaction that causal

	Fact contradicted…		
	…statement 1	…statement 2	…statement 3
Causal scenarios			
Did event in statement 1 occur?	--	20	10
Did event in statement 2 occur?	**85**	--	30
Did event in statement 3 occur?	**80**	**80**	--
Control scenarios			
Did event in statement 1 occur?	--	0	5
Did event in statement 2 occur?	5	--	0
Did event in statement 3 occur?	5	20	--

Table 2. The percentages of "no" answers in Experiment 2 to questions about the occurrence of the events in the first, second, or third statement, depending on whether the scenario was causal or control, and on whether the facts contradicted the first, second, or third statement in the scenario. The bold cells highlight domino effects.

simulation predicts: for causal scenarios, the participants rejected statements more often when they followed a contradicted statement than when they preceded one, but the difference did not hold for control scenarios (Wilcoxon test, z = 3.28, p = .001, Cliff's δ = .71). Planned comparisons elucidated the interaction: the numbers underneath the diagonals are reliably larger than those above for causal scenarios (Wilcoxon test, z = 3.70, p = .0002, Cliff's δ = .79) but not for control scenarios (Wilcoxon test, z = 1.65, p = .10, Cliff's δ = .20). The domino effect did not decline reliably when the fact contradicted the first statement, i.e., 85% of participants rejected the second statement and 80% of participants rejected the third statement (Wilcoxon test, z = 1.00, p = .32).

The participants judged whether or not an event occurred. But, refutations should also affect the subjective probabilities of the events. To extend the present results, Experiment 3 therefore used the same design, but its participants had to estimate the likelihood of the events in the scenarios.

4. Experiment 3

Experiment 3 used the same materials and design as Experiment 2, but the participants estimated the likelihood of events. Given, say, the following problem:

Robert heard a creak in the hall closet.

The faucet dripped.

The lawn sprinklers started.

In fact, Robert did not hear a creak in the hall closet.

they evaluated the likelihood that a particular event, such as the faucet dripped, on a seven-point Likert scale, ranging from *very unlikely* to *very likely*. On separate trials,

they made these judgments but only about those events that the facts did not contradict, e.g.:

How likely is it that the lawn sprinklers started?

Likelihood estimates reflect subjective probabilities, and so they allow participants to make a more refined response than merely accepting or rejecting an event. Under the assumption that participants should rarely hold beliefs with complete certainty, the scale prohibited responses of complete certainty (i.e., a probability of 1.0) or impossibility (i.e., a probability of 0.0).

4.1 Method

4.1.1 Participants. 24 participants (8 males, mean age = 31.8 years) from the same population as before completed the study for monetary compensation.

4.1.2 Design and procedure. The study used the same design and procedure as Experiment 2, but it presented participants with a different task. Participants made their evaluations of statements on a Likert scale that ranged from 1 (very unlikely) to 7 (very likely) through a midpoint of 4.

4.2 Results and discussion

Table 3 presents the participants' mean estimates of likelihood. They rated causal statements as less likely than control statements (M_{causal} = 3.36 vs $M_{control}$ = 5.52; Wilcoxon test, z = 3.86, p = .0001, Cliff's δ = .67). Their mean estimates of likelihood were 4.10, 4.35, and 4.87 when the first, second, and third statement was contradicted,

respectively (Page's trend test, $z = 3.53$, $p = .0004$). They likewise rated statements as less likely when they followed a contradicted statement than when they preceded one, but only reliably for causal scenarios; the interaction was reliable (Wilcoxon test, $z = 3.13$, $p = .001$, Cliff's $\delta = .51$). Planned comparisons elucidated it: the numbers in cells below the diagonal in Table 3 were reliably lower than the numbers above the diagonal for causal scenarios (Wilcoxon test, $z = 3.15$, $p = .002$, Cliff's $\delta = .52$) but not for control scenarios (Wilcoxon test, $z = .77$, $p = .44$, Cliff's $\delta = .10$). The domino effect did not dissipate reliably when the fact contradicted the first statement: the difference between participants' mean ratings of likelihood was not reliable (2.83 and 2.54 for the second and third statements, respectively, Wilcoxon test, $z = .60$, $p = .55$, Cliff's $\delta = .12$).

Experiment 3, and the studies that preceded it, show that causal scenarios are susceptible to domino effects. However, as a reviewer observed, it may be the case that other sorts of sequences yield domino effects as well – the previous studies did not

	Fact contradicted…		
	…statement 1	…statement 2	…statement 3
Causal scenarios			
How likely is it that event in statement 1 occurred?	--	4.33	4.42
How likely is it that event in statement 2 occurred?	**2.83**	--	3.96
How likely is it that event in statement 3 occurred?	**2.54**	**2.08**	--
Control scenarios			
How likely is it that event in statement 1 occurred?	--	5.50	5.75
How likely is it that event in statement 2 occurred?	5.58	--	5.38
How likely is it that event in statement 3 occurred?	5.46	5.50	--

Table 3. Mean estimates of likelihood for the first, second, and third statements in Experiment 3, depending on whether the fact contradicted the first, second, or third statement in a scenario. Estimates of likelihood ranged from 1 (very unlikely) to 7 (very likely). The bold cells highlight domino effects.

examine the issue. Participants may not have construed the control problems in each study as a temporal sequence – the problems could have been interpreted as a set of unrelated events. Hence, Experiment 4 modified the control problems to ensure that they presented temporal sequences to participants.

5. Experiment 4

Experiment 4 was identical to Experiment 3 in every respect – it used the same materials, design, and task as the previous study – except that it modified the control problems. In the previous studies, the control problems were of the following sort:

Robert heard a creak in the hall closet.

The faucet dripped.

The lawn sprinklers started.

Such descriptions could have been interpreted as temporal sequences or as a set of independent, unrelated events with no clear temporal relations. In Experiment 4, the control problems were modified so that they described a clear temporal sequence, e.g.:

Robert heard a creak in the hall closet.

Then the faucet dripped.

Then the lawn sprinklers started.

The causal scenarios were similarly modified. If participants exhibited domino effects for temporal sequences, it follows that domino effects are relevant for many sorts of sequence, not just causal ones. Otherwise, domino effects are a phenomenon specific to causal reasoning.

5.1 Method

5.1.1 Participants. 50 participants (25 females, mean age = 37.3 years) from the same population as before completed the study for monetary compensation.

5.1.2 Design and procedure. The study used the same design and procedure as Experiment 3, but it modified the materials so that the second and third statements in the descriptions were prefaced with the temporal adverb, "Then".

5.2 Results and discussion

Table 4 presents the participants' mean likelihood estimates, which shows that they pattern similarly to the previous study. Participants rated the likelihood of causal statements lower than temporal statements (M_{causal} = 3.92 vs $M_{temporal}$ = 5.42; Wilcoxon test, $z = 4.80$, $p < .0001$, Cliff's $\delta = .64$). When the first, second, or third statement was contradicted, their estimates of likelihood were 4.24, 4.55, and 5.26, respectively (Page's trend test, $z = 4.75$, $p < .0001$). For causal scenarios, statements that followed contradicted statement were rated less likely than those that preceded contradicted statements; the interaction was reliable (Wilcoxon test, $z = 4.53$, $p < .0001$, Cliff's $\delta = .54$). And planned comparisons further revealed the domino effect: for causal scenarios, the numbers in cells below the diagonal in Table 4 were reliably lower than the numbers above the diagonal (Wilcoxon test, $z = 4.95$, $p < .0001$, Cliff's $\delta = .61$); the effect was not reliable for temporal scenarios (Wilcoxon test, $z = 1.57$, $p = .12$, Cliff's $\delta = .13$). When the fact contradicted the first statement, the domino effect did not dissipate: participants

	Fact contradicted…		
	…statement1	…statement 2	…statement 3
Causal scenarios			
How likely is it that event in statement 1 occurred?	--	4.68	5.67
How likely is it that event in statement 2 occurred?	**2.95**	--	4.34
How likely is it that event in statement 3 occurred?	**2.97**	**2.84**	--
Control scenarios			
How likely is it that event in statement 1 occurred?	--	5.68	5.68
How likely is it that event in statement 2 occurred?	5.43	--	5.32
How likely is it that event in statement 3 occurred?	5.46	4.98	--

Table 4. Mean estimates of likelihood for the first, second, and third statements in Experiment 4, depending on whether the fact contradicted the first, second, or third statement in a scenario. Estimates of likelihood ranged from 1 (very unlikely) to 7 (very likely). The bold cells highlight domino effects.

rated the second and third statements as equally unlikely (2.96 and 2.98 for the second and third statements, respectively, Wilcoxon test, $z = .32$, $p = .75$, Cliff's $\delta = .01$).

If the participants had been minimalists then they should have judged all the statements that they evaluated as highly likely – the rejection of one statement should have minimal effect on the interpretation of the scenario. Instead, their judgments of likelihood reflected domino effects for the causal sequences, but not for the temporal sequences. The results of Experiment 4, as well as the results of Experiments 1-3, suggest that domino effects are a robust phenomenon of causal reasoning. And all four experiments corroborated the causal simulation hypothesis.

6. General discussion

The theory of causal simulation predicts that when a fact contradicts an event in a causal scenario, it should initiate a domino effect, i.e., the retraction of subsequent events

in the chain. Four studies validated its occurrence in causal scenarios. The participants understood the task, because they rejected statements that the facts directly contradicted (Experiment 1). Consider, say, the following problem:

Tony pressed the accelerator.

The car lurched forward.

The fender slammed into a tree.

In fact, Tony did not press the accelerator.

Most of the participants no longer believed that Tony pressed the accelerator. The small minority of participants who persisted in the belief may have done so out of confusion or carelessness. Nevertheless, all the participants tended to reject events subsequent to the contradiction of the link in the causal chain. No domino effect occurred, however, for control scenarios of independent events that should not elicit causal simulations. In a subsequent study (Experiment 2), participants no longer had to assess a statement that the facts contradicted to ensure that this unnecessary task did not confuse them; the domino effect still occurred for causal scenarios, but not for control scenarios. The effect occurred when the task required the participants to rate the subjective probability of statements (Experiment 3), and when the materials made explicit that the control scenario described a temporal sequence (Experiment 4).

All four experiments presented participants with an assertion that contradicted an event in a causal sequence. Other studies have examined how people use their causal knowledge to infer counterfactual scenarios. Rips's (2010) studies, for instance, presented participants with descriptions of the components of fictitious devices. They then answered counterfactual conditional questions of the sort, "If component C had not

operated; would component A have operated?" for devices in which component A caused component C to function. But, neither this study nor other studies of counterfactual conditionals (e.g., Sloman & Lagnado, 2005) used materials designed to reveal whether people exhibit domino effects. Nevertheless, the theory predicts that domino effects should occur with counterfactuals, as in the following problem:

David put a book on the shelf.

The shelf collapsed.

The vase broke.

If David had not put a book on the shelf, would the vase have broken?

They should respond, "no." Of course, a counterfactual possibility that satisfies the *if*-clause of the question above is one in which David puts an anvil on the shelf, not a book. Hence, minimalism predicts that they have no reason to respond, "no". Yet, reasoners should flout minimalism whether they reason about contradictions or counterfactuals.

One limitation of the present studies is that statements in the scenarios were so few and so simple. We used such problems to ensure that participants noticed the contradictions, and that they would use "bridging" inferences (Clark, 1975) to establish causal simulations. Previous studies had shown that participants often lost track of contradictions (Otero & Kintsch, 1992). No reason exists, however, to doubt that domino effects should occur in studies using more natural descriptions. Of course, after a very long causal chain, or one that is difficult to simulate (see, e.g., Johnson & Ahn, 2015), the effects may start to dissipate.

Another limitation of the studies, albeit a deliberate one, is that the order of events in the causal simulations was the same as the order of the statements in the scenarios, e.g.:

Harry pulled the trigger.

The gun fired.

The bullet shattered a window.

The causal chain leads from pulling the trigger, to the firing of the gun, to the shattering of the window. The sequence is the same as the order of the statements. The following description is analogous, but it uses the past perfect to make clear the temporal order of events:

The bullet shattered a window.

The gun had fired.

Harry had pulled the trigger.

Causal simulation still predicts that domino effects should occur in such scenarios, though the effect may be attenuated, because the causal chain may be less obvious.

The domino effect challenges the well-known doctrine of "minimalism" in the revision of beliefs, i.e., when individuals discover that a fact conflicts with their beliefs, they should make a minimal revision to their beliefs in order to accommodate the new fact (e.g., Gärdenfors, 1988; Harman, 1986; James, 1907; Levi, 1991; Quine, 1992). Minimalism, by definition, puts a premium on the preservation of knowledge. It therefore predicts that reasoners should never reject information unless it is necessary to do so in order to preserve consistency. Apart from unique causes, such as vitamin C deficiency

causing scurvy, it is never necessary to retract the effect of a cause, because it is always possible that some other cause brought it about. So, the domino effect runs counter to minimalism, and reflects the idea that intuitive simulations of causal relations embody strong causation in which the stated cause had no alternative (e.g., Goldvarg & Johnson-Laird, 2001; Khemlani, Barbey, & Johnson-Laird, 2014).

Perhaps minimalism can be salvaged: theorists could construct a model of minimalism that accommodates bridging inferences, enthymemes, presuppositions, and other such considerations. But, proponents of minimalism have avoided such a radical step, because doing so makes it almost impossible to count up the number of changes to propositions. Indeed, a notion of minimalism that accommodates implied content may be chimerical: there does not seem to be an effectively computable way of assessing that any change is minimal. Likewise, a more accommodating theory of minimalism would need to explain why people rate explanations that violate minimalism – explanations that introduce content that did not occur in the premises – as more plausible and probable than minimal ones (e.g., Johnson-Laird et al., 2004; Khemlani & Johnson-Laird, 2011, 2012, 2013; Legrenzi & Johnson-Laird, 2005). Perhaps the biggest difficulty for minimalism is exactly the phenomenon that our experiments establish: individuals disbelieve more than the minimum of a description needed to accommodate a conflicting fact. No obvious account compatible with minimalism appears to explain this violation. So, we pass the burden of formulating a new account of minimalism to its proponents.

Could minimalism be a normative principle rather than a description of human evaluations? That is, it could characterize the ideal way in which reasoners and reasoning systems ought to work. Violations of its constraint are therefore violations of

rationality. As such, minimalism is embodied in many systems in artificial intelligence (e.g., Euzenat, 2015). Likewise, recent psychological accounts of causal reasoning suggest that people maintain a bias towards simplicity (Lombrozo, 2007); such a bias predicts both that minimal causal inferences are rational, and that reasoners should prefer them to more complex causal inferences (but cf. Zemla, Sloman, Bechlivanidis, & Lagnado, 2017). In our view, two different issues are at stake here. On the one hand, simple explanations are easier to formulate and to work with. A long tradition exists in science in favor of Occam's razor: the avoidance of multiplying entities unnecessarily. If two theories make the same prediction, it is rational to prefer the simpler one. On the other hand, the occurrence of a contradiction between the facts of the matter and a causal hypothesis is a sign that the hypothesis is wrong. The problem calls for diagnosis. And, in daily life, diagnosis is often a basis for a decision about what to do. When causal systems go wrong, the premium on diagnosis is that it should be accurate. If the starter on your car does not turn over, it could be because the battery is dead, there is a short in the circuit, or the starter is broken. A simple explanation refers to just one of these defects, but an accurate explanation may call for more than one of them – and in many contexts, reasoners appear to flout simplicity (Johnson, Jin, & Keil, 2014; Johnson, Valenti, & Keil, 2019; Khemlani, Sussman, & Oppenheimer, 2011; Zemla et al., 2017). Explanatory accuracy is critical for rationality, not minimalism.

The hypothesis of causal simulation postulates that individuals use their knowledge of causal relations to simulate events. Various phenomena follow as a consequence (see, e.g., Khemlani & Johnson-Laird, 2011, 2012, 2013). The knowledge that books can be heavy and shelves flimsy enables individuals to simulate the situation

in which the weight of a book causes a shelf to break. Reasoners appear to prefer a *complete* causal simulation from the initial cause to the final effect in the scenario, and in which each effect has a cause (Johnson-Laird et al., 2004; Korman & Khemlani, 2018, under review; Zemla et al., 2017). On learning that the book was not put on the shelf, reasoners can halt the simulation, and, as a consequence, cease to believe that the subsequent events occurred in the chain. In our studies, the participants made much more than minimal changes. And they did so in a systematic and predictable way. In particular, they sought to drop the causal consequences of events that facts had denied, and one such modification led to another like dominos toppling. They ceased to believe a whole sequence of events, not just the minimal one that the statement of a fact contradicted. This phenomenon, in turn, implies that they aimed to incorporate the facts according to their background knowledge of causal relations.

In conclusion, individuals make causal simulations. So, when facts contradict a proposition, a domino effect occurs in a causal scenario but not in one describing independent events that elicit no causal links. All four of our experiments corroborated this hypothesis. We conclude that causal simulations take precedence over minimalism in coping with the consequences of contradictions.

Acknowledgements

This research was supported by a grant from the Naval Research Laboratory to SK to study causal reasoning and by NSF Grant No. SES 0844851 to PJL to study deductive and probabilistic reasoning. We are grateful to Ruth Byrne, Monica Bucciarelli, Sam

Glucksberg, Adele Goldberg, Geoff Goodwin, Sam Johnson, Zach Horne, Joanna Korman, and Marco Ragni, for their helpful comments and criticisms.

References

Baltag, A., Gierasimczuk, N., & Smets, S. (2011). Belief revision as a truth-tracking process. In *Proceedings of the 13th Conference on Theoretical Aspects of Rationality and Knowledge* (pp. 187-190). ACM.

Clark, H. H. (1975). Bridging. In R. C. Schank & B. L. Nash-Webber (Eds.), *Theoretical issues in natural language processing*. New York: Association for Computing Machinery.

Coste-Marquis, S., Konieczny, S., Mailly, J. G., & Marquis, P. (2014). On the revision of argumentation systems: Minimal change of arguments statuses. In *Proceedings of the 14th International Conference on Principle of Knowledge Representation and Reasoning* (pp. 52-61).

Euzenat, J. (2015). Revision in networks of ontologies. *Artificial Intelligence*, 228, 195-216.

Evans, J. S. B. T. (2012). Questions and challenges for the new psychology of reasoning. *Thinking & Reasoning, 18,* 5-31.

Fermé, E., & Hansson, S. O. (2011). AGM 25 years. *Journal of Philosophical Logic*, 40, 295-331.

Gärdenfors, P. (1982). Epistemic importance and minimal changes of belief. *Australasian Journal of Philosophy, 62,* 136-157.

Gärdenfors, P. (1988). *Knowledge in flux: Modeling the dynamics of epistemic states.* Cambridge, MA: MIT Press.

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

Harman, G. H. (1986). *Change in view: Principles of reasoning.* Bradford, MA: Bradford Books.

James, W. (1907). *Pragmatism – A new name for some old ways of thinking*. New York: Longmans.

Johnson, S. G., & Ahn, W. K. (2015). Causal networks or causal islands? The representation of mechanisms and the transitivity of causal judgment. *Cognitive Science*, 39, 1468-1503.

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

Johnson, S. G. B., Valenti, J. J., & Keil, F. C. (2019). Simplicity and complexity preferences in causal explanation: An opponent heuristic account. *Cognitive Psychology,* 113, 101222.

Johnson-Laird, P. N., Girotto, V., & Legrenzi, P. (2004). Reasoning from inconsistency to consistency. *Psychological Review, 111,* 640-661.

Khemlani, S., Barbey, A.K., & Johnson-Laird, P.N. (2014). Causal reasoning: mental computations, and brain mechanisms. *Frontiers in Human Neuroscience*, *8*, 1-15.

Khemlani, S., & Johnson-Laird, P.N. (2011). The need to explain. *Quarterly Journal of Experimental Psychology*, *64*, 276-88.

Khemlani, S., & Johnson-Laird, P.N. (2012). Hidden conflicts: Explanations make inconsistencies harder to detect. *Acta Psychologica, 139,* 486–491.

Khemlani, S., & Johnson-Laird, P. N. (2013). Cognitive changes from explanations. *Journal of Cognitive Psychology, 25,* 139–146.

Khemlani, S., Mackiewicz, R., Bucciarelli, M., & Johnson-Laird, P.N. (2013). Kinematic mental simulations in abduction and deduction. *Proceedings of the National Academy of Sciences*, *110 (42)*, 16766–16771.

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

Korman, J., & Khemlani, S. (2018). How people detect incomplete explanations. In C. Kalish, M. Rau, T. Rogers, & J. Zhu (Eds.), Proceedings of the 40th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society.

Korman, J., & Khemlani, S. (under review). Explanatory completeness. Manuscript under review at *Acta Psychologica*.

Legrenzi, P., & Johnson-Laird, P. N. (2005). The evaluation of diagnostic explanations for inconsistencies. *Psychologica Belgica, 45,* 19-28.

Levi, I. (1991). *The fixation of belief and its undoing*. Cambridge, MA: Cambridge University Press.

Oaksford, M., & Chater, N. (2013). Dynamic inference and everyday conditional reasoning in the new paradigm. *Thinking & Reasoning, 19,* 346-379.

Otero, J. & Kintsch, W. (1992). Failures to detect contradictions in a text: What readers believe versus what they read. *Psychological Science, 3,* 229-235.

Over, D. E. (2009). New paradigm psychology of reasoning. Thinking & Reasoning 15(4): 431.

Quine, W. V. O. (1992). *Pursuit of truth.* Cambridge, MA: Harvard University Press.

Rips, L. (2010). Two causal theories of counterfactual conditionals. *Cognitive Science, 34,* 175-221.

Sloman, S., & Lagnado, D. (2005). Do we "do"? *Cognitive Science, 29,* 5-39.

Walsh, C., & Johnson-Laird, P.N. (2009). Changing your mind. *Memory & Cognition, 37*, 624-631.

Zemla, J. C., Sloman, S., Bechlivanidis, C., & Lagnado, D. A. (2017). Evaluating everyday explanations. *Psychonomic Bulletin & Review*, 1-13.

Appendix

Table A1. The materials used in Experiments 1-4. The first, second, and third statements were provided to participants, and one of them was contradicted. In Experiment 4, the second and third statements were preceded by the adverb, "Then…".

First statement	Second statement	Third statement
Causal		
David put a book on the shelf	The shelf collapsed	The vase broke
Sammy pushed a button on his cell phone	The phone dialed	The answering machine began
Fred threw a water balloon at George	The balloon hit him	He was drenching wet
Harry pulled the trigger	The gun fired	The bullet shattered a window
Sarah turned on the kitchen light	The bulb burst	Glass fell on the kitchen counter
Tony pressed the accelerator	The car lurched forward	The fender slammed into a tree
Control		
Molly laughed at a man on TV	The dog barked	The door opened
Robert heard a creak in the hall closet	The faucet dripped	The lawn sprinklers started
Frank gave a telephone book to Ron	The clock struck five	Ron was running late
Peter mowed the lawn	The mailman arrived	The dog chased a car
Katie switched off the washing machine	The cat meowed	The children came home from school
Walter lit a candle	The radio was blaring	The delivery man rang the doorbell

Supplementary Materials

Supplementary materials, including raw data, analysis scripts, code for the experiments, and materials, are available at https://osf.io/zwnjs/.