The Quarterly Journal of Experimental Psychology

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/pqje20

The need to explain

Sangeet S. Khemlani [a] & Philip N. Johnson-Laird [a]

[a] Department of Psychology, Princeton University, Princeton, NJ, USA

Available online: 06 Jun 2011

To cite this article: Sangeet S. Khemlani & Philip N. Johnson-Laird (2011): The need to explain, The Quarterly Journal of Experimental Psychology, DOI:10.1080/17470218.2011.592593

To link to this article: http://dx.doi.org/10.1080/17470218.2011.592593

THE QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY
i*First*, 1–13

Psychology Press
Taylor & Francis Group

The need to explain

Sangeet S. Khemlani and Philip N. Johnson-Laird
Department of Psychology, Princeton University, Princeton, NJ, USA

How do reasoners deal with inconsistencies? James (1907) believed that the rational solution is to revise your beliefs and to do so in a minimal way. We propose an alternative: You *explain* the origins of an inconsistency, which has the side effect of a revision to your beliefs. This hypothesis predicts that individuals should spontaneously create explanations of inconsistencies rather than refute one of the assertions and that they should rate explanations as more probable than refutations. A pilot study showed that participants spontaneously explain inconsistencies when they are asked what follows from inconsistent premises. In three subsequent experiments, participants were asked to compare explanations of inconsistencies against minimal refutations of the inconsistent premises. In Experiment 1, participants chose which conclusion was most probable; in Experiment 2 they rank ordered the conclusions based on their probability; and in Experiment 3 they estimated the mean probability of the conclusions' occurrence. In all three studies, participants rated explanations as more probable than refutations. The results imply that individuals create explanations to resolve an inconsistency and that these explanations lead to changes in belief. Changes in belief are therefore of secondary importance to the primary goal of explanation.

Keywords: Inconsistency; Explanations; Belief revision; Minimalism; Reasoning.

In a memorable scene in *All The President's Men*, Bernstein and Woodward's account of the Watergate investigation, Woodward decides to meet Deep Throat, his secret source inside the Nixon administration. They arrange to meet in a secluded garage in the middle of the night, but Deep Throat does not show up. The reporters recount Woodward's thoughts:

> Woodward was becoming worried. Deep Throat rarely missed an appointment. In the dark, cold garage, Woodward began thinking the unthinkable. It would not have been difficult for Haldeman to learn that the reporters were making inquiries about him. Maybe Deep Throat had been spotted? Woodward followed? People crazy enough to hire Gordon Liddy and Howard Hunt were crazy enough to do other things. Woodward got mad at himself for becoming irrational. (Bernstein & Woodward, 1974, p. 172)

Far from being irrational, however, the reporter's reasoning reflects a struggle to reconcile conflicting information. The process of reasoning about inconsistencies is likely to lead individuals to abandon

Correspondence should be addressed to Sangeet Khemlani, Department of Psychology, Princeton University, Princeton, NJ, 08540, USA. E-mail: khemlani@princeton.edu

This research was supported by a National Science Foundation Graduate Research Fellowship to the first author and by National Science Foundation Grant SES 0844851 to the second author to study deductive and probabilistic reasoning. We are grateful for helpful criticisms from Jeremy Boyd, John Darley, Sam Glucksberg, Adele Goldberg, Geoffrey Goodwin, Matt Johnson, Olivia Kang, Niklas Kunze, Max Lotstein, and Laura Suttle.

http://www.psypress.com/qjep
DOI:10.1080/17470218.2011.592593

some of their conclusions and even perhaps some of their premises.

From the standpoint of orthodox logic, the choice of which information to change or reject is an arbitrary one (see Jeffrey, 1981), and computer scientists have proposed separate logics to handle inconsistencies (for a review, see Brewka, Dix, & Konolige, 1997). Psychologists have attempted to uncover systematic patterns in these choices (Dieussaert, Schaeken, De Neys, & d'Ydewalle, 2000; Elio & Pelletier, 1997; Johnson-Laird, Girotto, & Legrenzi, 2004; Politzer & Carles, 2001; Rehder & Hastie, 1996; Revlis, Lipkin, & Hayes, 1971; Schlottmann & Anderson, 1995; Walsh & Johnson-Laird, 2009). So far, they have been unable to formulate clear-cut procedures for determining which premises to abandon in the face of inconsistency. However, theorists from William James (1907) onwards have argued that changes to propositions should be as minimal as possible (Gärdenfors, 1988; Harman, 1986; Levi, 1991; Quine, 1992). As James (1907, p. 59) wrote: "[The new fact] preserves the older stock of truths with a minimum of modification, stretching them just enough to make them admit the novelty." The idea, which is alternatively known as the "maxim of minimum mutilation" (Quine, 1992, p. 14), the "principle of conservatism" (Harman, 1986, p. 46) and the "criterion of informational economy" (Gärdenfors, 1982, p. 136), posits that individuals should modify, add, or retract as little information as possible—a view we refer to simply as *minimalism*.

Minimalism requires a tractable procedure for evaluating the amount of change a revision makes to a set of propositions and for choosing which of equally minimal changes to make. Harman (1986) argues that an appropriate metric is to "take the sum of the number of (explicit) new beliefs added plus the number of (explicit) old beliefs given up" (p. 61). Such a measure is simple, and it is difficult to specify an alternative measure that is both effective and testable. Indeed, other theorists have proposed similar ways to count changes (Elio & Pelletier, 1997; Hiddleston, 2005), and so we adopt this metric too. As Elio and Pelletier point out, minimalism implies that changes to specific beliefs, such as, *Pat received a heavy blow to the head*, are more minimal than changes to generalizations, such as, *If a person receives a heavy blow to the head then that person forgets some preceding events*. They explain that "for classical belief revision theories, the intuition driving the idea of entrenching [if P then Q] over other types of sentences is not because material implication per se is important, but because 'lawlike relations' are often expressed in sentences of this form" (Elio & Pelletier, 1997, p. 427).

Minimalism is a hypothesis about how to change your beliefs in the face of inconsistency, and it presupposes that such revisions are your primary psychological goal in coping with an inconsistency. In our view, however, the presupposition has no warrant. In daily life, when an inconsistency arises because a fact collides with the consequences of your beliefs, your primary goal is to understand how the inconsistency could have occurred in the first place, because its origins are likely to have consequences for how you should act. Consider the conflict between Woodward's expectations and Deep Throat's failure to show: It is no mere fluke that Woodward attempts to resolve the inconsistency by considering an explanation—it bears directly on whether he should get out of the garage as fast as possible. Despite his worry that he is being irrational, the process of reasoning to the best explanation is a hallmark of rationality (Harman, 1965, 1986), because it is a prerequisite for sensible action. A mere revision to beliefs, whether minimal or not, is not so useful a guide. This intuition underlies Craik's case for the construction of models of the world in order to conclude which is the best course of action: "If the organism carries a 'small-scale model' of external reality and of its own possible actions within its head, it is able to . . . react to future situations before they arise, utilize the knowledge of past events in dealing with the present and the future, and in every way to react in a much fuller, safer, and more competent manner to the emergencies which face it" (Craik, 1943, p. 61).

An alternative to the idea of belief revision is accordingly the *explanatory* hypothesis, which

postulates that the first goal in coping with an inconsistency is to explain its origin. A plausible explanation is likely to imply changes to beliefs (Johnson-Laird et al., 2004; Thagard, 1989). These changes may, or may not, be minimal, depending on the nature of the explanation. But, an explanation is a novel proposition that introduces new entities, properties, and relations over and above those giving rise to the inconsistency. The explanatory hypothesis and minimalism accordingly yield different predictions about how individuals deal with inconsistencies.

Consider the following illustrative example:

1. If a person pulls the trigger then the pistol fires.

   Someone pulled the trigger but the pistol did not fire.

   What follows?

Individuals detect the inconsistency (Johnson-Laird et al., 2004). And minimalism, as we pointed out earlier, implies that a change to a categorical proposition, such as the second assertion in the example, is more minimal than a change to a generalization, such as the first assertion (see Elio & Pelletier, 1997; Harman, 1986, pp. 59–63). In fact, individuals tend to revise the conditional when they are asked explicitly which assertion they would give up (Elio & Pelletier, 1997). Various explanations for this phenomenon exist, including syntactic ones (Politzer & Carles, 2001) and semantic ones based on mental models (Johnson-Laird et al., 2004). But, an alternative hypothesis is that any plausible explanation is likely to place the onus of doubt on the general claim embodied in the conditional, because individuals are familiar with the idea that certain conditions can "disable" causal claims (see, e.g., Cummins, 1995), such as the one in Problem 1. Hence, a crucial question is whether inconsistencies trigger explanations.

The present studies sought to establish whether reasoners spontaneously create explanations that resolve inconsistencies or instead revise the assertions giving rise to them, perhaps in a minimal way. A pilot study accordingly presented participants with inconsistencies similar to Problem 1 above, and their task was a neutral one with respect to the prediction. They had to state what followed from the assertions. The answers generated from the pilot study were used to construct materials for Experiments 1, 2, and 3. Participants in these experiments evaluated four sorts of conclusions given premises similar to Problem 1. Two sorts of conclusion introduced new information in order to explain the inconsistency, for example:

a. The safety had not been taken off the pistol.
b. The person is scared of guns and refuses to touch them.

   The first explanation has the indirect effect of refuting the generalization expressed in the if–then premise, whereas the second has the indirect effect of refuting the categorical assertion that someone pulled the trigger. The other sorts of conclusion have these effects directly:

c. The pistol doesn't always fire if the trigger is pulled.
d. The trigger of the gun was not actually pulled.

The experiments used various procedures in which reasoners rated the respective probabilities of explanations (a and b above) and minimal revisions (c and d above).

PILOT STUDY

To generate plausible explanations and refutations for subsequent experiments, a pilot study examined reasoners' spontaneous responses when faced with inconsistent scenarios. They read scenarios, such as:

2. If a person is bitten by a viper then the person dies.

   Someone was bitten by a viper, but did not die.

   What follows?

Such scenarios are judged to be inconsistent when individuals are asked to evaluate them (Johnson-Laird et al., 2004), but the experiment was designed not to draw attention to the inconsistency, but rather to elicit the participants' spontaneous reactions by posing the question, "what follows?".

Their task was to draw any conclusion that came to mind, and we examined the implications of their conclusions for the explanatory hypothesis. The task of determining in an objective way whether or not a conclusion is an explanation is difficult, because theorists disagree about what counts as an explanation—a matter that has generated its own considerable literature in the philosophy of science (see, e.g., Salmon, 1989). One influential view is that an explanation logically entails the explanandum—that is, whatever stands in need of explanation (Hempel, 1965), but this view is plainly not directly applicable in the present case because the explanandum is a logical inconsistency. We therefore adopted two simple criteria for an explanation. First, explanations of events, such as those in Problem 2, call for the introduction of a new proposition over and above what is stated in the problem. Second, this new proposition provides a causal account of the final outcome, such as the survival of the person bitten by the viper. Such an explanation should have the side effect that it rules out, or at least modifies, one of the other assertions in the inconsistent set. Examples of such conclusions to the problem above that introduce a new proposition about an entity, or a property, or a relation between entities, and that explain the outcome, are as follows:

The person got prompt medical attention.

The person wore heavy protective gloves.

The person spent years building up immunity to viper venom.

The criteria for conclusions to count as explanations were accordingly that they were propositions that introduced novel entities, properties, or relations, which did not occur in the set of inconsistent assertions, and that they provided a causal account of the final outcome (at the likely expense of overruling one of the assertions). The criteria are necessary conditions for almost all explanations, and, as a matter of fact, they appeared to be sufficient in our results. As far as we could tell, such cases in the experiment had the intuitive force of explanations, and this claim certainly held for the most frequent conclusions (see Appendix).

Method

Participants

A total of 29 participants completed the study for monetary compensation on Mechanical Turk, an online platform hosted on Amazon.com (for a discussion on the validity of results from this platform, see Paolacci, Chandler, & Ipeirotis, 2010). None of the participants had received any training in logic. Online participants were chosen so as to allow conclusions to be generalized to a wider population than those typically tested in a university.

Design, materials, and procedure

Participants carried out problems based on inconsistencies that arose from two assertions of the grammatical form, *If A then B; A, but not B*. They were told to suppose that the assertions were true, and their task was to respond to the question, "What follows?" Each participant responded freely to 12 problems, which were drawn from five different domains: biology, economics, mechanics, psychology, and natural phenomena. The appendix (columns 1 and 2) presents the two assertions in each of the problems. The generalizations were all highly plausible and similar to those used by Johnson-Laird et al. (2004). The study was administered using an interface written in PHP, HTML, and Javascript. Participants were invited to type their responses into a text box provided on the screen.

Results and discussion

The key contrast was whether the participants' conclusions implied a direct refutation of at least one of the premises, or else spontaneously went beyond them to add new entities, properties, or relations, in a putative explanation of the inconsistency. The appendix illustrates the contrast in examples of the main sorts of conclusion. Two research assistants with no knowledge of the hypotheses under consideration served as independent raters. They decided (a) whether each conclusion refuted the generalization or the categorical premise, and (b) whether or not each conclusion added information to the premises. (They agreed on 80% of trials, Kendall's $W = .74$, $p < .0001$, and

reconciled their differences through discussion.) When a conclusion refuted both the generalization and the categorical assertion, it counted as a refutation of the categorical (contrary to the explanatory hypothesis). The results showed that the participants added new information in 69% of their conclusions and thereby refuted at least one premise, but they directly refuted a premise in only 31% of their conclusions (Wilcoxon test, $z = 3.15$, $p < .005$). The value of Cliff's δ (a nonparametric effect size indicator whose value ranges from –1 to 1; see Cliff, 1993) was .67. Likewise, 24 of the 29 participants added information more often than not (binomial test, $p < .001$).

The participants drew the following percentages of conclusions:

Explanation refuting the generalization	65%
Explanation refuting the categorical assertion	4%
Direct refutation of the generalization	24%
Direct refutation of the categorical assertion	7%

The conclusions implied that the generalizations were false on 89% of trials and implied that the categorical assertions were false on 11% of trials (Wilcoxon test, $z = 4.65$, $p < .00001$, Cliff's $\delta = .98$), and 28 out of 29 participants produced conclusions refuting generalizations more often than not (binomial test, $p < .00001$). The tendency for explanations to refute generalizations was greater than the tendency for direct refutations to refute generalizations, and this interaction was reliable (Wilcoxon test, $z = 3.67$, $p < .0005$, Cliff's $\delta = .63$). Direct refutations occurred on nearly a third of the trials, and so the pragmatics of the task did not prevent the participants from making them.

The participants tended to infer explanations of the inconsistencies rather than to make revisions to the premises. Elio (1998) alluded to the role of explanation in revising beliefs, but no previous study has shown that logically untrained individuals tend to *explain* inconsistencies rather than to revise the inconsistent propositions. The next three experiments bore out the results of the pilot study. Its results also corroborated the preference for the refutation of generalizations rather than the refutations of categorical assertions—a result that is contrary to minimalism but that has been observed many times (Elio & Pelletier, 1997; Dieussaert et al., 2000; Politzer & Carles, 2001; Walsh & Johnson-Laird, 2009).

EXPERIMENTS 1, 2, AND 3

Granted that individuals tend to resolve inconsistencies by explaining their origins rather than by revising the offending assertions, they should judge such explanations as more probable than revisions. Experiments 1, 2, and 3 tested this prediction in a variety of ways. To examine estimates of the probabilities of explanations and revisions to premises, we devised materials based on the most frequent conclusions that the participants drew in the pilot study (see Appendix), and which all met the two criteria in the case of explanations. For instance, for the scenario:

3. If a person does regular aerobic exercises then that person strengthens his or her heart.

   Someone did regular aerobic exercises but did not strengthen his heart.

The four categories of conclusion were:

a. An explanation with the consequence of refuting the generalization: The person had a congenital heart defect.
b. An explanation with the consequence of refuting the categorical: The person was too busy during the workweek.
c. A direct refutation of the generalization: Aerobic exercises do not always strengthen your heart.
d. A direct refutation of the categorical: The person did not do the exercises regularly.

Method

Participants

All three experiments were carried out online using the Amazon.com platform described earlier; 21 participants volunteered for Experiment 1, 17 for Experiment 2, and 25 for Experiment 3.

Design, materials, and procedure

Participants carried out problems based on inconsistencies between a generalization (*If A then B*) and a categorical assertion (*A, but not B*). For each problem, they were instructed to suppose that the two assertions were true and were given four sorts of conclusion (see Appendix, columns 3–6). The experiments examined estimates of probabilities in three different ways, but in each case the participants carried out every condition and provided estimates of the probabilities of the four sorts of conclusion. Experiment 1 presented them as four separate options, and the participants chose the most probable one. Experiment 2 used the same options, but the task was to rank order them in terms of their probabilities from 1 (*the most probable*) to 4 (*the least probable)*. Experiment 3 used a more naturalistic procedure in which the participants estimated the probability of one option on a percentage scale on each trial, but the experiment as a whole yielded an overall comparison of the probabilities of the four sorts of conclusion. In all other respects, the procedures were similar to the one used for the pilot study.

Results and discussion

Table 1 summarizes the results of the three experiments. They all yielded the same principal phenomenon. In Experiment 1, the participants chose explanations as the most probable conclusion on 64% of trials and direct refutations on 36% of trials (Wilcoxon test, $z = 2.34$, $p < .05$, Cliff's $\delta = .56$). Consistent with the explanatory hypothesis and previous observations (Elio & Pelletier, 1997), participants rated conclusions that refuted generalizations, either with explanations or directly, as more probable than refutations of categoricals (Wilcoxon test, $z = 2.7$, $p < .01$, Cliff's $\delta = .68$). However, the interaction was reliable: For explanations, the participants tended to rate those that refuted generalizations as more probable than those that refuted categoricals, but this trend was reversed for direct refutations (Wilcoxon test, $z = 4.02$, $p < .0001$, Cliff's $\delta = .98$). The interaction shows that the experimental procedure did not inhibit participants from making direct refutations. In Experiment 2, the participants yielded a reliable trend in their rankings (Kendall's $W = .68$, $p < .0001$), and explanations that refuted generalizations had the highest ranked probability. Likewise, in Experiment 3, the estimates of probabilities also yielded a reliable trend (Kendall's $W = .19$, $p < .005$), and once again explanations that refuted generalizations had the highest ranked probabilities. Ratings for explanations that refuted generalizations were higher than those that refuted categoricals, but there was no difference between ratings of the two sorts of direct refutation (Wilcoxon test, $z = 3.46$, $p < .001$, Cliff's $\delta = .63$). The relative probabilities of the different conclusions were not entirely consistent from one version of the experiment to another. In all three versions, however, explanations that refuted generalizations were rated as most probable, but direct refutations of categoricals were second most probable in Experiments 1 and 3, but ranked the least probable in Experiment 2. This discrepancy probably reflected the role of direct comparisons between conclusions in the different tasks. The ranking task in Experiment 2 forced the participants to consider all four sorts of conclusion on each trial, whereas the other two experiments allowed the participants to focus on individual conclusions.

Table 1. *The results of Experiments 1, 2, and 3*

	Experiment		
Type of conclusion	*1*	*2*	*3*
Explanations refuting generalizations	58	1.7	70
Explanations refuting categoricals	6	2.9	44
Direct refutations of generalizations	7	2.3	59
Direct refutations of categoricals	29	3.0	61

Note: The percentages with which the participants chose each of the four sorts of conclusion as most probable in Experiment 1. The mean in which the participants rank ordered the four sorts of assertion in Experiment 2 (rank 1 = most probable). The participants' mean estimates of probabilities (expressed in percentages) in Experiment 3.

GENERAL DISCUSSION

Theories of belief revision posit that people resolve inconsistencies by revising or abandoning their beliefs. The minimalist hypothesis predicts that people should abandon categoricals more often than generalizations, because they seek to make as few changes as possible to their information. In contrast, our hypothesis is that individuals are more concerned to formulate explanations that resolve inconsistencies than to revise the conflicting propositions. They make explanations first, and these explanations then imply changes to their beliefs. Because explanations are propositions that add new entities, properties, and relations to those in the premises, individuals are prepared to sacrifice minimal change in order to achieve their explanatory goal.

In fact, most individuals propose explanations that indirectly refute generalizations and that are far from minimal changes. Such explanations are often what psychologists refer to as "disabling conditions", which provide cases in which the generalization fails. Individuals tend to refrain from inferences from generalizations with salient disabling conditions (e.g., Byrne, 1989; Cummins, 1995). Because of their propensity to envisage disabling conditions, their explanations are indeed more likely to invoke such conditions than to imply that a proposition about a specific individual or entity is wrong. The pilot study corroborated this pattern, and a reliably smaller proportion of conclusions were direct refutations of one or other of the assertions. Likewise, participants tended to select such explanations as the most probable (in Experiment 1), to evaluate them as having the highest rank of probability (Experiment 2), and to assign them the highest probability (Experiment 3). These studies corroborate the finding that people are prepared to make nonminimal changes to resolve inconsistencies (Dieussaert et al., 2000; Elio & Pelletier, 1997; Politzer & Carles, 2001; Walsh & Johnson-Laird, 2009). The studies also supported the explanatory hypothesis: The principal goal that most individuals have in resolving an inconsistency is to explain its origins. Their explanations then imply revisions to the inconsistent assertions.

The studies we report have several limitations. First, the conditional premises are causal generalizations about which reasoners have background knowledge. Explanations of inconsistencies about causal generalizations may be easier to construct. However, explanations in daily life often have an underlying causal structure (for a review, see Keil, 2006), and so we sought to include inconsistencies that are relevant to everyday experiences. Another limitation is that the generalizations were all expressed in conditional assertions. In a separate pilot study, we examined the effects of different types of generalizations by comparing conditionals (e.g., "If a person eats this dish then that person gets indigestion") with universals (e.g., "All people who eat this dish get indigestion") and generics (e.g., "People who eat this dish get indigestion"). We found similar patterns of response—namely, that participants tended to generate explanations that refuted generalizations (on 90% of trials).

Could participants' tendency to explain inconsistencies reflect a process that does not normally occur in daily life? Reasoners in the pilot study were instructed to suppose that the information presented to them was true, and such an instruction might have suggested that they should not reject the generalizations. Yet, they did reject them. The task was to answer the question, "what follows?", which is neutral with respect either to offering an explanation or to making a minimal change. But, could this question somehow have inhibited direct refutations in the pilot study? It is hard to see why such an inhibition would have occurred, and, as we mentioned earlier, a substantial minority of conclusions did make such refutations. The question does not even arise for the subsequent studies, which used various methods for the participants to evaluate the probabilities of different sorts of conclusion. The robust finding with all these methods was that explanations, especially those that indirectly refuted generalizations, were judged to have the highest probability of all. The greater probability assigned to explanations that refute a premise rather than to direct refutations of the premise is an instance of the "conjunction" fallacy in which a conjunction is in error judged to be more probable than its constituents (Tversky & Kahneman,

1983). Hence, the consensus among our participants suggests that the tasks were reasonably representative of everyday thinking.

The results were inconsistent with minimalism. Contrary to what standard psychological theories of belief revision have assumed (e.g., James, 1907), people react to an inconsistency, not by revising their beliefs, but by seeking an explanation that resolves the inconsistency. Any revision in beliefs appears to be a by-product of this explanatory process. Proponents of minimalism might argue that the explanations that participants generated were in fact minimal changes with respect to their background beliefs. Explanations do rely on knowledge, and so they can hardly undermine the knowledge on which they depend. One problem with this defence of minimalism is to test it, because of the difficulty in assessing changes to tacit beliefs. What we can say, however, is that a simple rejection of a categorical premise in our studies does not call for any change to tacit beliefs whatsoever. Hence, even the appeal to background beliefs does not save minimalism (for additional evidence against this appeal, see Walsh & Johnson-Laird, 2009). Of course, our results leave open the possibility that minimalism is a normative theory (see Harman, 1986, p. 7), in which case, they show that untrained individuals depart from a canon of rationality.

In our view, a rational response to an inconsistency is to formulate an explanation that resolves the inconsistency. The reason is that explanations provide a better guide to future action than do revisions in belief (Craik, 1943; Johnson-Laird, 2006; Keil, 2006) and are central to the way we communicate our understanding of the world (Chi, de Leeuw, Chiu, & LaVancher, 1994; Lombrozo, 2007). To revert to our opening example, if Woodward explains Deep Throat's failure to make their appointment in terms of foul play, then the explanation suggests a quite different course of action than one, say, based on a traffic jam in Washington, DC. Cognitive scientists have accordingly begun to examine how explanations guide learning and judgement (Ahn, Marsh, Luhmann, & Lee, 2002; Murphy & Allopenna, 1994), foster conceptual development (de Leeuw & Chi, 2003; Murphy, 2000), and facilitate exploration (Legare, in press). And the present results suggest that explanatory reasoning is fundamental to resolving inconsistencies (see also Johnson-Laird et al., 2004; Khemlani & Johnson-Laird, 2010; Legare, Gelman, & Wellman, 2010).

How do individuals create explanations? According to the theory of mental models, the fundamental unit of causal explanations is a chain consisting of a cause and its effect (Johnson-Laird, 2006), and so individuals seek such explanations in order to resolve inconsistencies. Since few empirical generalizations of the sort shown in the Appendix are universal—that is, they all have exceptions—a plausible manoeuvre is to use knowledge to construct a causal scenario that explains the inconsistency by yielding a counterexample to the generalization—that is, a disabling condition. A computer program illustrates the process for examples akin to the one above (see Johnson-Laird et al., 2004):

5. If the trigger is pulled then the pistol will fire.
   The trigger is pulled.
   But, the pistol does not fire.

The program constructs a model of the possibility described in the first two assertions:

trigger pulled pistol fires

The fact that the pistol did *not* fire is inconsistent with this model. Nevertheless, the conditional expresses a useful idealization, and so the program treats it as the basis for the mental models shown in Table 2, Models A. In its knowledge-base, the program has fully explicit models of various ways in which a pistol may fail to fire—that is, disabling conditions, such as, if the pistol doesn't have any bullets in it, if it is damaged, or if its safety catch is on. The model of the facts above triggers one of these sets of models corresponding, say, to the first of these cases, and the relevant model modulates the facts to create a model of the explanation (see Table 2, Model B). The new proposition in this model, not(bullets), can in turn trigger a causal antecedent from another set of models in

Table 2. *A model of the facts and the counterfactual possibility, and a model of a possible explanation*

				Representation of
Models A:		trigger pulled	not(pistol fires)	the facts
		trigger pulled	pistol fires	a counterfactual possibility
			…	
Model B:	not(bullets)	trigger pulled	not(pistol fires)	an explanation

the knowledge base representing a cause for the absence of bullets in a pistol—for example, if a person empties the bullets from the pistol. The resulting possibility explains the inconsistency: A person emptied the pistol, and so it had no bullets. And the counterfactual possibilities above yield the claim: If the person hadn't emptied the pistol then it would have had bullets, and it would have fired. In sum, the fact that the pistol did not fire has been used to create an explanation from knowledge, which in turn refutes the generalization and transforms it into a counterfactual claim (Byrne, 2005).

Mental models are accordingly a viable way in which reasoners might represent explanations, and they account for why individuals prefer explanations that refute generalizations to those that refute categorical statements (Johnson-Laird et al., 2004). However, other theories may also be compatible with the explanatory hypothesis. Oaksford and Chater (2010) provide a treatment of generalizations as defeasible assertions, and they can in principle accommodate participants' preferences. Likewise, other theorists emphasize the role of defeasibility and uncertainty in interpreting generalizations (e.g., Pfeifer & Kleiter, 2011). Most (if not all) assertions in daily life are defeasible, including the premises we provided to participants in the present experiments. Indeed, these studies suggest a mechanism for defeasibility: Reasoners first interpret assertions as generalizations without exceptions, and they then reinterpret those assertions based on explanations they construct. This mechanism explains why reasoners are able to detect inconsistencies in the first place (Khemlani & Johnson-Laird, 2010), and it also explains their explanatory preferences.

In conclusion, the natural way in which individuals are likely to deal with an inconsistency is not to edit the inconsistent propositions—minimally or otherwise—to restore consistency, but rather to seek an explanation that resolves the anomaly. The inconsistencies in our scenarios arose from a clash between, on the one hand, a generalization and a categorical assertion and, on the other hand, an incontrovertible fact. The scenarios were from five distinct empirical domains—biology, economics, mechanics, psychology, and natural phenomena—and so explanations to resolve them were causal. They tended to be counterexamples to the generalizations in the scenarios, which were idealizations rather than ironclad claims.

Original manuscript received 21 December 2010
Accepted revision received 27 February 2011
First published online day month year

REFERENCES

Ahn, W., Marsh, J., Luhmann, C., & Lee, K. (2002). Effect of theory-based feature correlations on typicality judgments. *Memory & Cognition*, *30,* 107–118.

Bernstein, C., & Woodward, R. (1974). *All the president's men*. New York, NY: Simon & Schuster.

Brewka, G., Dix, J., & Konolige, K. (1997). *Nonmonotonic reasoning: An overview*. Stanford, CA: CLSI Publications, Stanford University.

Byrne, R. M. J. (1989). Everyday reasoning with conditional sequences. *Quarterly Journal of Experimental Psychology*, *41A,* 141–166.

Byrne, R. M. J. (2005). *The rational imagination*. Cambridge, MA: MIT Press.

Chi, M. T. H., de Leeuw, N., Chiu, M. H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18,* 439–477.

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, *114,* 494–509.

Craik, K. (1943). *The nature of explanation*. Cambridge, MA: Cambridge University Press.

Cummins, D. D. (1995). Naive theories and causal deduction. *Memory & Cognition*, *23,* 646–658.

de Leeuw, N., & Chi, M. T. H. (2003). The role of self-explanation in conceptual change learning. In G. Sinatra & P. Pintrich (Eds.), *Intentional conceptual change* (pp. 55–78). Mahwah, NJ: Lawrence Erlbaum Associates.

Dieussaert, K., Schaeken, W., De Neys, W., & d'Ydewalle, G. (2000). Initial belief state as a predictor of belief revision. *Current Psychology of Cognition*, *19,* 277–288.

Elio, R. (1998). How to disbelieve p → q: Resolving contradictions. In M. A. Gernsbacher & S. J. Derry (Eds.), *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates.

Elio, R., & Pelletier, F. J. (1997). Belief change as propositional update. *Cognitive Science*, *21,* 419–460.

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.

Harman, G. H. (1965). The inference to the best explanation. *Philosophical Review*, *74,* 88–95.

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

Hempel, C. (1965). *Aspects of scientific explanation and other essays in the philosophy of science*. New York, NY: Free Press.

Hiddleston, E. (2005). A causal theory of counterfactuals. *Nous*, *39,* 632–657.

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

Jeffrey, R. (1981). *Formal logic: Its scope and limits*. New York, NY: McGraw Hill.

Johnson-Laird, P. N. (2006). *How we reason*. Oxford, UK: Oxford University Press.

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

Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology*, *57,* 227–254.

Khemlani, S., & Johnson-Laird, P. N. (2010). Explanations make inconsistencies harder to detect. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Legare, C. (in press). Exploring explanation: Explaining inconsistent information guides hypothesis-testing behavior in young children. *Child Development*.

Legare, C. H., Gelman, S. A., & Wellman, H. M. (2010). Inconsistency with prior knowledge triggers children's causal explanatory reasoning. *Child Development*, *81,* 929–944.

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

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

Murphy, G. L. (2000). Explanatory concepts. In R. A. Wilson & F. C. Keil (Eds.), *Explanation and cognition* (pp. 361–392). Cambridge, MA: MIT Press.

Murphy, G. L., & Allopenna, P. D. (1994). The locus of knowledge effects in concept learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *20,* 904–919.

Oaksford, M., & Chater, N. (Eds.). (2010). Conditionals and constraint satisfaction: Reconciling mental models and the probabilistic approach? *Cognition and conditionals: Probability and logic in human thinking* (pp. 309–334). Oxford, UK: Oxford University Press.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5,* 411–419.

Pfeifer, N., & Kleiter, G. D. (2011). Uncertain deductive reasoning. In K. Manktelow, D. E. Over, & S. Elqayam (Eds.), *The science of reasoning: A Festschrift for Jonathan St B. T. Evans* (pp. 145–166). Hove, UK: Psychology Press.

Politzer, G., & Carles, L. (2001). Belief revision and uncertain reasoning. *Thinking & Reasoning*, *7,* 217–234.

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

Rehder, B., & Hastie, R. (1996). The moderating influence and variability on belief revision. *Psychonomic Bulletin & Review*, *3,* 499–503.

Revlis, R., Lipkin, S. G., & Hayes, J. R. (1971). The importance of universal quantifiers in a hypothetical reasoning task. *Journal of Verbal Learning and Verbal Behavior*, *10,* 86–91.

Salmon, W. (1989). *Four decades of scientific explanation*. Minneapolis, MN: University of Minnesota Press.

Schlottmann, A., & Anderson, N. H. (1995). Belief revision in children: Serial judgment in social cognition and decision-making domains. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21,* 1349–1364.

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences*, *12,* 435–467.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90,* 292–315.

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

APPENDIX

The problems used in the experiments

Table A1. *The 12 problems used and examples of the most frequent responses from the pilot study*

		Explanation		*Refutation*	
Generalization	*Categorical*	*Refutes generalization*	*Refutes categorical*	*Of generalization*	*Of categorical*
If a person is bitten by a viper then they die	Someone was bitten by a viper but did not die	The person received an antidote	The person was wearing heavy clothing	A viper's bite is not always deadly	The person was not bitten by a viper
If a person does regular aerobic exercises then that person strengthens his or her heart	Someone did regular aerobic exercises but did not strengthen his or her heart	The person had a congenital heart defect	The person was too busy during the workweek	Aerobic exercises do not always strengthen your heart	The person did not do the exercises regularly
If a car's engine is tuned in the special way then its fuel consumption goes down	This car's engine was tuned in the special way but its fuel consumption did not go down	The car had engine problems that increased consumption	The driver accidentally read the wrong gauge[a]	Fuel consumption doesn't always go down with special tuning	Fuel consumption actually did go down
If graphite rods are inserted into a nuclear reactor, then its activity slows down	Graphite rods were inserted into this nuclear reactor but its activity did not slow down	The graphite rods were incorrectly inserted in the reactor	Aluminium rods were inserted in the nuclear reactor instead	Reactor activity does not always slow down when graphite rods are inserted	Graphite rods were not inserted into the nuclear reactor
If the aperture on a camera is narrowed, then less light falls on the film	The aperture on this camera was narrowed but less light did not fall on the film	It was completely dark, so there was no light at all	The mechanism controlling the aperture was broken	Less light doesn't always fall on the film with a narrowed aperture	The aperture of the camera was not narrowed
If a person pulls the trigger then the pistol fires	Someone pulled the trigger but the pistol did not fire	The safety had not been taken off the pistol	The person is scared of guns and refuses to touch them	The pistol doesn't always fire if the trigger is pulled	The trigger of the gun was not actually pulled
If a substance such as butter is heated then it melts	This piece of butter was heated but it did not melt	The heat was too low to melt the butter	The substance was actually hard wax	Substances like butter don't always melt when heated	The substance was not actually butter
If these two substances come into contact with one another then there is an explosion	These two substances came into contact with one another but there was no explosion	There was not enough of either of the substances	The substances repelled each other at the last moment	Contact between these substances doesn't always cause an explosion	The substances did not actually come into contact

If someone is very kind then he or she is liked by others	Someone was very kind but was not liked by others	Sometimes excessive kindness comes off insincere or condescending	The others are reserved and find it hard to show that they like someone	Not all people who are very kind are liked by others	The very kind person actually was liked by the others
If a person receives a heavy blow to the head then that person forgets some preceding events	Pat received a heavy blow to the head but did not forget any preceding events	Pat was wearing a helmet at the time	The blow actually glanced off his temple	Heavy blows to the head don't always cause lost memories	Pat did not receive a heavy blow to the head
If people make too much noise at a party then the neighbours complain	People made too much noise at a party but the neighbours did not complain	The neighbours were away on summer vacation	The neighbours notified the police the next morning	The neighbours don't always complain about loud parties	People did not make very much noise at the party
If the banks cut interest rates then the economy increases	The banks cut interest rates but the economy did not increase	Cutting rates is not enough in an economic decline	The banks changed their decision at the last minute	The economy doesn't always increase if banks cut interest rates	The banks did not actually cut interest rates

Note: The problems used in the experiments were generalizations combined with categorical assertions. Responses were explanations that refute either the generalization or the categorical assertion, and direct refutations of the generalization or the categorical assertion (as judged by two independent raters).

[a]This particular item is ambiguous due to a lack of specificity, as it could be understood as refuting the categorical (e.g., if the fuel consumption did go down but the driver read the wrong gauge) or refuting the generalization (e.g., if the gauge indicated that the tuning was performed properly when in fact it was not). We reran the analyses without the item and observed no differences in the statistical results.