Explanations make inconsistencies harder to detect

Sangeet Khemlani and P.N. Johnson-Laird
{khemlani, phil}@princeton.edu
Department of Psychology
Princeton University
Princeton, NJ 08540 USA

Abstract

What role do explanations play in reasoning about inconsistencies? We postulate that when people create explanations, they use them to resolve conflicting information. This hypothesis predicts that inconsistencies should be harder to detect once individuals have in mind an explanation of the inconsistency. We report four experiments that tested this prediction. Experiments 1a and 1b corroborated the effect when participants made inferences from inconsistent assertions. Experiment 2 compared the effect of explanations of inconsistencies with those of a similarly demanding task. Experiment 3 ruled out a potential confound.

Keywords: inconsistency, explanations, belief revision, reasoning, principle of resolution

Introduction

The word 'why' is used to elicit explanations for the mysteries of daily life. Why is my car making that noise? Why didn't the Redskins win last Sunday? Why isn't my experiment working? Indeed, a central feature of human rationality is the ability to construct explanations of observed behaviors and phenomena (Harman, 1965). Recent research has explored the function and developmental trajectory of explanatory reasoning (Keil, 2006; Wellman, Hickling, & Schult, 1997). There is consensus among researchers that explanations are related to causal inference (Johnson-Laird, Girotto, & Legrenzi, 2004; Sloman, 2005; Walsh & Johnson-Laird, 2009), and that explanations impact reasoning, categorization, and learning (Lombrozo, 2006). Less is known about the contexts under which explanations are generated, i.e., it is unclear when and how individuals decide to produce explanations.

How do you reveal what a person understands about some subject matter? One way is to ask the person to explain it, because explanations require individuals to communicate their knowledge and beliefs about the phenomenon in question. Explanations can also occur in other tasks that draw upon general knowledge. For instance, explanations are useful when you are learning new information (Amsterlaw & Wellman, 2006; Chi, De Leeuw, Chiu, & Lavancher, 1994; Crowley & Siegler, 1999; Rittle-Johnson, 2006), and they help to predict future behaviors (Anderson & Ross, 1980; Einhorn & Hogarth, 1986; Lombrozo & Carey, 2006; Ross, Lepper, Strack, & Steinmetz, 1977). Individuals spontaneously produce explanations when they try to form categories (Shafto & Coley, 2003) and when they judge how well concepts cohere with one another (Murphy & Medin, 1985; Palatano, Chin-Parker, & Ross, 2006). We propose that an additional function of explanatory reasoning is to resolve inconsistencies.

Explanations resolve inconsistencies

Consider the following:

If people are tired then they go to sleep.
A person was tired, but he did not go to sleep.

The two assertions are inconsistent, i.e., they cannot both be true. Given such an inconsistency, it is felicitous to ask: "why not?" But the same question is infelicitous when the assertions are obviously consistent with one another:

If people are tired then they go to sleep.
A person was not tired, and he did not go to sleep.

It seems strange to elicit an explanation for consistent assertions, and reasoners are likely to balk at such a request. Thus, an inconsistency calls for people to search for explanations, while an explanation is less appropriate when expectations are met.

We hypothesize that individuals resolve a set of inconsistent causal assertions by using an explanation to interpret each assertion, a view we call the *principle of resolution*. The principle assumes that when an inconsistency is detected among a set of assertions, reasoners construct explanations to restore consistency to the set (Johnson-Laird et al., 2004). They then interpret the assertions based on the consequences of the explanations. Consider the inconsistency above. One explanation for the person not going to sleep is that he was under some deadline, and so pursued his work despite his fatigue. The explanation provides an exception to the generalization that if people are tired they go to sleep. However, instead of abandoning it, reasoners are likely to construe it as an idealization that holds by default: it is true in many cases, but tolerates exceptions. The assertion may be interpreted as something akin to the generic assertion, i.e., 'people who are tired go to sleep' (Khemlani, Leslie, Glucksberg, & Rubio-Fernandez, 2007; Leslie, 2008). The principle of resolution thus allows individuals to use explanations to resolve inconsistencies by weakening the initial interpretation to that of an idealization rather than a universal truth.

One potential side effect of the principle is that when reasoners have an explanation of an inconsistency in mind, they may overlook the inconsistency on subsequent assessments of the assertions. If they interpret the conditional as an idealization, their new interpretation may

prevent them from detecting the conflict between the two assertions. Indeed, they may even forget that the reason for constructing the explanation in the first place was to resolve an inconsistency. To test this prediction, participants in four experiments were asked to detect an inconsistency after they had carried out various tasks.

Experiments 1a and 1b

Experiments 1a and 1b examined whether reasoners spontaneously construct explanations when faced with inconsistent scenarios, and whether those explanations made it more difficult to detect inconsistencies. They were presented with problems such as:

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

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

The participants in Experiment 1a evaluated the consistency of two assertions, either before or after they stated what follows from the assertions. The participants in Experiment 1b evaluated the consistency of the assertions before or after they responded to the question, "why not?" When individuals make an inference from inconsistent assertions, they should tend to infer explanations. The principle of resolution posits that when people create explanations, they interpret the assertions in the light of their explanation. It predicts an interaction: when individuals create an explanation first, they should be less accurate subsequently at detecting inconsistencies in comparison with those who have not created an explanation.

Method

Participants. 36 participants were recruited for Experiment 1a, and 40 participants were recruited for Experiment 1b. They volunteered through an online platform hosted through Amazon.com, and they completed the study for monetary compensation. None of the participants had received any training in logic.

Design and Procedure. On each trial, participants were given a set of consistent or inconsistent assertions (see Appendix A). Half of the problems presented a generalization (1) that was inconsistent with a categorical assertion (2), e.g.,

1. *If someone is very kind then he or she is liked by others.*
2. *Someone was very kind but was not liked by others.*

For the remaining problems, the inconsistency was eliminated by dropping the first clause in the categorical assertion (4), e.g.,

3. *If someone is very kind then he or she is liked by others.*
4. *Someone was not liked by others.*

Participants received an equal number of consistent and inconsistent problems, and carried out two tasks in succession for each problem, a consistency task and a task designed to elicit explanations. For the consistency task, participants had to answer the question, "Can both of these statements be true at the same time?" They responded by pressing one of two buttons marked "Yes" or "No". In Experiment 1a, participants also performed an inferential task, i.e., they answered the question, "What, if anything, follows from the statements above?" In Experiment 1b, they performed a more orthodox explanation task, i.e., they answered the question, "Why not?" They typed their responses into a text box provided on the screen. They were unable to see their response to the first task when they carried out the second task. In Experiment 1a, 20 participants performed the inferential task before the consistency task, and 16 participants performed the two tasks in the opposite order. In Experiment 1b, 20 participants performed the explanation task before the consistency task, and 20 performed the two tasks in the opposite order. All of the problems were similar to the two examples above, and participants received each set of contents only once. Each participant received the problems in a different random order.

Results and Discussion

Table 1 reports the proportions of trials on which participants correctly evaluated the assertions as consistent or inconsistent in Experiment 1a. Overall, participants were more accurate on consistent problems than inconsistent problems (77% vs. 50%, Wilcoxon test, $z = 3.27$, $p < .005$, Cliff's $d = .42$), and the group that carried out the consistency task first was marginally more accurate than the group that initially made an inference about the assertions (70% vs. 58%, Mann-Whitney test, $z = 1.66$, $p = .10$, Cliff's $d = .32$). These main effects were a consequence of the low rate of accuracy on inconsistent problems observed for the group that carried out the inferential task first. Their responses corroborated the principle of resolution, and the predicted interaction was significant: the group that initially carried out the inferential task was less accurate at detecting inconsistencies than consistencies, while the group that initially carried out the consistency task was just as accurate at detecting either type of problem (Mann-Whitney test, $z = 3.03$, $p < .005$, Cliff's $d = .59$). Accuracy in the evaluation of consistency in Experiment 1a therefore depended on whether or not participants initially made an inference about the assertions. The effect is likely to reflect the use of inferences that explain the inconsistency.

Table 1: The percentages of correct evaluations of consistency in Experiment 1a depending on whether participants carried out the evaluation or the inferential task first.

	Inconsistent problems	Consistent problems
Group that carried out the consistency task first	73	68
Group that carried out the inferential task first	33	84

Table 2 reports the proportions of correct responses in Experiment 1b. Participants were far more accurate at detecting consistencies than inconsistencies (89% vs. 45%, Wilcoxon test, $z = 4.00$, $p < .0001$, Cliff's $d = .69$). The group that initially evaluated the consistency of the assertions was more accurate than the group that initially provided an explanation (79% vs. 56%, Mann-Whitney test, $z = 3.07$, $p < .005$, Cliff's $d = .66$). And the predicted interaction was significant: the difference between accuracies on inconsistent vs. consistent problems was greater for the group that carried out the explanatory task first (Mann-Whitney test, $z = 2.02$, $p < .025$, Cliff's $d = .48$). As in Experiment 1a, participants in Experiment 1b were less accurate at detecting inconsistencies when they initially provided an explanation.

These results support the principle of resolution, which predicted that explanations would make it more difficult to detect inconsistencies. However, it is possible that the difficulty to detect inconsistencies could have occurred because the explanation and inferential tasks were inherently more difficult. In other words, there may not have been anything unique about the explanation task, and the same effects could have been observed had reasoners performed any task that increased processing load. The evidence for such an account is mixed: in Experiment 1a, participants who initially made an inference were more accurate at detecting consistencies than participants who initially carried out the consistency task (84% vs. 68%).

Table 2: The percentages of correct evaluations of consistency in Experiment 1b depending on whether participants carried out the evaluation or the explanation task first.

	Inconsistent problems	Consistent problems
Group that carried out the consistency task first	64	93
Group that carried out the explanation task first	27	86

Hence, a difference in processing load cannot readily explain this pattern of results. It should have decreased performance on both sorts of problem, but in fact the participants did better on the consistent problems. In contrast, a difference in processing load could explain the results of Experiment 1b, because in this case the participants who answered the question 'why not?' first, went on to evaluate the consistency of both sorts of problem worse than those participants who began with this evaluation task. Experiment 2 therefore sought to determine whether any demanding task could dull reasoners' sensitivity to inconsistencies, or whether explanations are unique in decreasing accuracy.

Experiment 2

To test whether explanations uniquely contribute to low rates of accuracy when individuals have to detect inconsistencies, the participants in this experiment evaluated the consistency of a set of assertions after carrying out one of two tasks: one group provided an explanation of the assertions and the other group decided whether some clauses of the assertions were more surprising than others. The surprisingness task was chosen because it required reasoners to take into account all the assertions, but it did not require them to construct explanations of inconsistencies. Those participants who performed the surprisingness task received trials such as the following one:

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

In light of these statements, which of the following is more surprising?

1. It's more surprising that the aperture on this camera was narrowed.

2. It's more surprising that less light did not fall on the film.

They received the same instructions for consistent trials, and responded by choosing between one of two alternative responses. Once their responses were registered, they carried out the consistency task. The other group of participants typed out their response to the question "Why not?" before completing the consistency task.

Method

Participants. 40 participants from the same online platform as in the previous studies and completed the experiment for monetary compensation.

Design and Procedure. Participants received an equal number of consistent and inconsistent problems, and received the same set of problems used in the previous study. Half the participants carried out the explanation task before the consistency task and the other half carried out the surprisingness task before the consistency task. They were unable to see their responses to the initial task when they carried out the consistency task. Participants received each set of contents only once, and each participant received the problems in a different randomized order.

Results and Discussion

Table 3 reports the proportions of correct responses in Experiment 2. The results again corroborated the principle of resolution. Participants were less accurate for inconsistent than consistent problems when they carried out the explanation task than when they carried out the surprisingness task (Mann-Whitney test, $z = 1.64$, $p = .05$, Cliff's $d = .30$). No decrease in accuracy was observed for consistent problems between the two groups (86% vs. 84%,

Mann-Whitney test, $z = .63$, $p = .53$). The results rule out the possibility that the effects reflected differences in processing load.

Table 3: The percentages of correct evaluations of consistency in Experiment 2 depending on whether participants carried out the surprisingness task first or the explanation task first.

	Inconsistent problems	Consistent problems
Group that carried out the surprisingness task first	75	86
Group that carried out the explanation task first	47	84

The experiment replicated the previous effect: participants who created explanations often went on to evaluate an inconsistent set of assertions as consistent, but the surprisingness task had no such effect. The study ruled out the possibility that any demanding mental task would yield the same results, because participants who rated how surprising the assertions were did not go on to err in their evaluation of the inconsistent problems. And both groups went on to evaluate consistent problems with no reliable difference in accuracy between them.

In Experiment 2, reasoners either carried out the surprisingness task or else the explanation task before judging the consistency of the assertions. That is, no participant was exposed to the two different task orders. Experiment 3 sought to extend the results to a context in which each participant carried out both tasks.

Experiment 3

Experiment 3 tested whether explanations impair evaluations of consistency more than judgments of surprisingness. On each trial, participants either provided an explanation, a judgment of surprisingness, or neither, before they evaluated the consistency of the assertions.

Method

Participants. 25 participants from the same online platform as in the previous studies completed the experiment for monetary compensation. None had received any training in logic.

Design and Procedure. Participants served as their own controls, and received an equal number of consistent and inconsistent problems. The materials consisted of those used in the previous studies. For a third of the trials, participants carried out only the consistency task; on another third, they carried out the surprisingness task before the consistency task; and on the remaining trials they carried out the explanation task before the consistency task. The three conditions were intermingled, and each participant received the problems in a different randomized order. Participants received each set of contents only once, and the contents were rotated over the three conditions so that each content occurred equally often in each condition in the experiment as a whole.

Results and Discussion

Table 4 provides the proportions of correct responses in Experiment 3. Participants were more accurate on consistent problems than inconsistent problems (71% vs. 52%, Wilcoxon test, $z = 2.38$, $p < .01$, Cliff's $d = .26$), and accuracy varied by the three types of trials (Friedman analysis of variance, $\chi^2 = 6.20$, $p < .05$). These main effects can be attributed to the drop in accuracy on inconsistent problems when participants had provided explanations.

The study yielded the predicted interaction between the type of trial and the consistency of the problem, i.e., participants were less accurate on inconsistent problems when they had carried out the explanation task than when they had carried out the surprisingness task or no prior task, whereas their accuracies for consistent problems were comparable to one another across the different tasks (Page's $L = 304.5$, $z = 2.55$, $p < .001$).

Table 4: The percentages of correct evaluations of consistency in Experiment 3 depending on whether participants carried out only the consistency task, the surprisingness task first, or the explanation task first.

	Inconsistent problems	Consistent problems
Consistency task only	60	70
Surprisingness task, then consistency task	56	76
Explanation task, then consistency task	40	68

As in the previous studies, Experiment 3 showed that explanations increased the likelihood that participants evaluated inconsistent assertions as consistent. The effect cannot be explained as a function of task demand, because participants did no better after they carried out the surprisingness task than after they had carried out no prior task. The study also extended the findings to a study in which the participants carried out all the different sorts of task. We conclude that the effect of explanations on consistency ratings is robust.

General Discussion

Across four experiments, participants erroneously evaluated inconsistent assertions as consistent after they had created an explanation for the inconsistency. Experiment 1a found that people produced the effect when they were asked to make inferences from the assertions, and Experiment 1b

extended the effect by directly eliciting explanations. Experiment 2 reproduced the effect by comparing those who formulated explanations with those who performed an unrelated task. Experiment 3 extended the effect to a context in which participants carried out the tasks in different orders. If participants had focused only on the assertions they were asked to read, the creation of an explanation should have had no effect on the evaluation of consistency in any of our experiments. Instead, the participants failed to detect inconsistencies as a result of creating explanations. When individuals resolve an inconsistency by explaining it, they are likely to establish a consistent interpretation of the facts of the matter and the original assertions. They have reasoned from inconsistency to consistency (see Johnson-Laird et al., 2004), and this newfound consistency makes it harder to detect the original inconsistency of the assertions.

Two gaps in the present account remain. First, the quality of the explanations that the participants created appeared to vary, but further research is need to interrelate this quality, say, to the latency of a correct evaluation of the inconsistent assertions. Second, the precise mechanism underlying the phenomenon has yet to be pinned down. When individuals explain an apparent inconsistency among a set of assertions, their explanation may sometimes rule out one of the assertions as false, and it may sometimes yield an idealized interpretation of a conditional generalization. For example, is the conditional assertion:

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

true or false? Given the further premise, say, that Viv was bitten by a viper, many people are likely to make the inference that Viv died. Yet, in answer to the preceding question, they might respond, "there are exceptions". In other words, the conditional expresses a truth that holds by default, i.e., a counterexample does not overturn it. In contrast, individuals are likely to judge that the conditional assertion:

If a person's brain is deprived of oxygen for 1 hour then the person dies.

is true unequivocally. And they might not be prepared to believe a description of an apparent counterexample.

The results of our experiments corroborate the principle of resolution, which states that when individuals detect an inconsistency, they formulate explanations to restore consistency. They subsequently can interpret the inconsistent assertions according to the consequences of their explanations. As a result they may treat conditional assertions as tolerating exceptions, which they can explain by invoking disabling conditions (Cummins, 1995). For example, consider the following problem:

If a person pulls the trigger then the pistol fires.

Someone pulled the trigger but the pistol did not fire.

If, like many of our participants, you explain the inconsistency by believing that there were no bullets in the pistol's chamber, then you have qualified the first assertion. It is true only when bullets are in the pistol's chamber, i.e., an enabling condition is satisfied. When bullets are not in the pistol's chamber, the conditional no longer hold (Johnson-Laird et al., 2004).

The present studies demonstrate the power and purpose of explanatory reasoning. Reasoners can draw inferences or answer the question 'why not?' without realizing that the set of assertions they reason about is inconsistent. The explanations they construct make it less likely that they will subsequently detect the inconsistency, because a plausible explanation serves to resolve the inconsistency. In some situations, this behavior is sensible and practical, because it allows individuals to revise their beliefs. In other situations, however, the behavior may account for striking lapses in reasoning. When a plausible explanation is available, regardless of whether it is true, reasoners may overlook glaring inconsistencies and behave in accordance with the explanation.

The present studies demonstrate the power and purpose of explanatory reasoning. Reasoners can draw inferences or answer the question 'why not?' without realizing that the set of assertions they reason about is inconsistent. The explanations they construct make it less likely that they will subsequently detect the inconsistency, because a plausible explanation serves to resolve the inconsistency. In some situations, this behavior is sensible and practical, because it allows individuals to revise their beliefs. In other situations, however, the behavior may account for striking lapses in reasoning. When a plausible explanation is available, regardless of whether it is true, reasoners may overlook glaring inconsistencies and behave in accordance with the explanation.

In sum, individuals who construct explanations of inconsistent assertions have difficulty evaluating those assertions as inconsistent. They do so erroneously, as the assertions remain in conflict with one another regardless of whether an explanation is available.

Acknowledgments

This research was supported by a National Science Foundation Graduate Research Fellowship to the first author, and by National Science Foundation Grant No. 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.

References

Amsterlaw, J., & Wellman, H.M. (2006). Theories of mind in transition: a microgenetic study of the development of false belief understanding. *Journal of Cognitive Development, 7,* 139-172.

Anderson, C.A., & Ross, L. (1980). Perseverence of social theories: The role of explanation in the persistence of discredited information. *Journal of Personality and Social Psychology, 39,* 1037-1049.

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

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

Crowley, K., & Siegler, R.S. (1999). Explanation and generalization in young children's strategy learning. *Child Development, 70,* 304-316.

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

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

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

Johnson-Laird, P.N. (2006). *How we reason.* Oxford: 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,* 225-254.

Khemlani, S., Leslie, S.J., Glucksberg, S., & Rubio-Fernandez, P. (2007). Do ducks lay eggs? How people interpret generic assertions. *In Proceedings of the 29th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Leslie, S.J. (2008). Generics: Cognition and acquisition. *Philosophical Review, 117,* 1-47.

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences, 10,* 464-470.

Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition, 99,* 167-204.

Murphy, G.L., & Medin, D.L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92,* 289-316.

Patalano, A. L., Chin-Parker, S., & Ross, B. H. (2006). The importance of being coherent: Category coherence, cross-classification, and reasoning. *Journal of Memory and Language, 54,* 407-424.

Rittle-Johnson, B. (2006). Promoting transfer: the effects of direct instruction and self-explanation. *Child Development, 77,* 1-15.

Ross, L., Lepper, M.R., Strack, F., Steinmetz, J. (1977). Social explanation and social expectation: Effects of real and hypothetical explanations on subjective likelihood. *Journal of Personality and Social Psychology, 35,* 817-829.

Shafto, P., & Coley, J.D. (2003). Development of categorization and reasoning in the natural world: novices to experts, naïve similarity to ecological knowledge. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 29,* 641-649.

Sloman, S.A. (2005). *Causal models: how we think about the world and its alternatives.* New York: Oxford UP.

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

Wellman, H.M., Hickling, A.K., & Schult, C.A. (1997). Young children's psychological, physical, and biological explanations. *New Directions for Child Development, 75,* 7-25.

Appendix A

The assertions used in the experiments (generalizations were paired with consistent or inconsistent categorical assertions).

Domain	Generalization	Consistent Categorical	Inconsistent Categorical
Biology/physiology	If a person is bitten by a viper then they die	Someone did not die	Someone was bitten by a viper but did not die
Biology/physiology	If a person does regular aerobic exercises then that person strengthens his or her heart	Someone did not strengthen his heart	Someone did regular aerobic exercises but did not strengthen his or her heart
Mechanical	If a car's engine is tuned in the special way then its fuel consumption goes down	This car's fuel consumption did not go down	This car's engine was tuned in the special way but its fuel consumption did not go down
Mechanical	If graphite rods are inserted into a nuclear reactor, then its activity slows down	The nuclear reactor's activity did not slow down	Graphite rods were inserted into this nuclear reactor but its activity did not slow down
Mechanical	If the aperture on a camera is narrowed, then less light falls on the film	Less light did not fall on the film	The aperture on this camera was narrowed but less light did not fall on the film
Mechanical	If a person pulls the trigger then the pistol fires	The pistol did not fire	Someone pulled the trigger but the pistol did not fire
Natural	If a substance such as butter is heated then it melts	This piece of butter did not melt	This piece of butter was heated but it did not melt
Natural	If these two substances come into contact with one another then there is an explosion	There was no explosion	These two substances came into contact with one another but there was no explosion
Psychological	If someone is very kind then he or she is liked by others	Someone was not liked by others	Someone was very kind but was not liked by others
Psychological	If a person receives a heavy blow to the head then that person forgets some preceding events	Pat did not forget any preceding events	Pat received a heavy blow to the head but did not forget any preceding events
Social/economical	If people make too much noise at a party then the neighbors complain	The neighbors did not complain	People made too much noise at a party but the neighbors did not complain
Social/economical	If the banks cut interest rates then the economy increases	The economy did not increase	The banks cut interest rates but the economy did not increase