DOI: http://dx.doi.org/10.5334/pb-45-1-19

Psychologica Belgica
2005, 45-1, 19-28.

THE EVALUATION OF DIAGNOSTIC EXPLANATIONS FOR INCONSISTENCIES

Paolo LEGRENZI
University IUAV and School of Advanced Studies in Venice, Italy

Philip N. JOHNSON-LAIRD
Princeton University, USA

When individuals detect an inconsistency between a fact and their beliefs, they revise their beliefs. They also use their causal knowledge to create explanations of what led to the inconsistency. According to the theory in the present paper, an ideal explanation is a chain of a cause and an effect, where the effect explains the inconsistency. Two experiments corroborated this account. When participants evaluated explanations for inconsistencies, they rated a conjunction of a cause and its effect as more probable than the cause alone, which they rated as more probable than the effect alone. This trend violates the laws of probability – it is an instance of the "conjunction fallacy". It also violates the common assumption that individuals make minimal changes to their beliefs.

> Today, almost 20 years after nonmonotonic reasoning was established as an important research topic, we have made considerable progress in the theoretical understanding of default reasoning. On the other hand, a satisfactory account of the computational properties of human commonsense reasoning still seems to be lacking.
>
> Brewka, Dix, and Konolige (1997)

Events in daily life sometimes surprise you, because they yield an inconsistency with what you believe. Suppose, for instance, that you believe the following propositions:

School of Advanced Studies Venice – Department of Psychology, Princeton University.

This research was supported by the School of Advanced Studies in Venice and by a grant to the second author from the National Science Foundation (Grant BCS 0076287) to investigate strategies in reasoning. We thank Professor Riccardo Varaldo, Director of the Scuola Superiore Sant'Anna of Pisa, for generously allowing us to carry out experiments at the school; and we thank Vittorio Girotto for help and advice.

Correspondence: Paolo Legrenzi, School of Advanced Studies Venice, D.D. 3488/U, 30123 Venice, Italy. E-mail: legrenzi@iuav.it; Philip N. Johnson-Laird, Department of Psychology, Princeton University, Princeton, NJ 08544, USA. E-mail: phil@princeton.edu

If the book was sent by an express service, then it will arrive in two days.
and:

The book was sent by an express service.

You infer validly: the book will arrive in two days. Two days go by, and then another two, with no sign of the book. You begin to wonder whether the book was sent express or whether, if it was, it will arrive in two days. At least one of these premises must be false, though logic cannot tell you which one it is. But, unlike systems of reasoning in artificial intelligence (see, e.g., Brewka et al., 1997), you don't just revise your beliefs. You try to envisage what is likely to have happened. You try to create a diagnostic explanation that resolves the inconsistency between your beliefs and the fact that the book has not arrived.

The resolution of inconsistencies depends on three main steps: you must detect an inconsistency, which is not always easy (see, e.g., Legrenzi, Girotto, & Johnson-Laird, 2003); you must revise your beliefs; and, perhaps as part of that process, you must try to explain the inconsistency. Such reasoning is almost always causal, and an ideal explanation is a chain from a cause to an effect that, in turn, resolves the inconsistency (see Johnson-Laird, Girotto, & Legrenzi, 2004). You construct such a chain from your available knowledge. And you may create several possible explanations. For example, you think: possibly, the wrong address was on the parcel and that caused it to go astray. According to the theory of mental models, a causal relation of this sort refers to a set of three possibilities (see Goldvarg & Johnson-Laird, 2001):

wrong address	astray
¬ wrong address	astray
¬ wrong address	¬ astray

Each line denotes a separate possibility, "¬" represents negation, and the consequent states cannot precede the antecedents in temporal order. It follows that it should be easier to infer an effect from its cause than to infer a cause from its effect. Given the cause (the wrong address), there is only one possibility and the effect (the parcel went astray) occurs in that possibility; whereas given the effect, as the preceding models show, there is more than one possibility and the cause does not occur in one of them. By consequence, it is possible to explain why the inference from cause to effect is indeed easier than the converse inference (Tversky & Kahneman, 1982).

When inconsistencies are of the form illustrated in our example, the model theory predicts that their explanations should tend to reject the conditional premise rather than the categorical premise (for more on rejection choice, see also Dieussaert, De Neys, & Schaeken, 2005; Revlin, Calvillo, & Ballard, 2005). Thus, the explanation that the wrong address was on the parcel is consistent with the categorical premise that the parcel was sent by an express service. But, it is inconsistent with the conditional claim that if so, it will arrive in two days. That claim holds only if such mishaps do not occur. A conditional, such as:

If the book was sent by an express service, then it will arrive in two days is compatible with three possibilities, granted that the book might arrive in two days if it was sent by some other means:

express service	arrive in two days
¬ express service	arrive in two days
¬ express service	¬ arrive in two days

Individuals are able to enumerate these possibilities, but they normally represent such a conditional with only two mental models:

express service	arrive in two days

. . .

where the ellipsis denotes a mental model with no explicit content. It is a place holder for the possibilities in which the antecedent of the conditional (the book was sent by an express service) is false (for corroboratory evidence, see Johnson-Laird & Byrne, 1991). In the example, the inconsistent fact (the book didn't arrive in two days) conflicts with the one explicit mental model of the conditional, and that is why explanations that resolve this inconsistency should tend to reject the conditional rather than the categorical premise.

The hypotheses about causal inferences and the rejection of conditionals yield a prediction about which, between the possible explanations resolving the inconsistency, was judged most probable. Consider again the putative explanation of our example:

1. The parcel had the wrong address and it went astray. (*Cause-and-effect*)

This explanation describes a cause and its effect, and the effect explains why the parcel hasn't arrived. It is an ideal explanation, and so individuals should rate it as having a higher probability than the cause alone:

2. The parcel had the wrong address. (*Cause*)

It is easier to infer the effect from the cause than to infer the cause from the effect. Hence, individuals should rate the cause alone as having a higher probability than the effect alone:

3. The parcel went astray. (*Effect*)

This predicted pattern of judgments is an instance of the "conjunction fallacy", in which a conjunction is judged to be more probable than its constituents (Tversky & Kahneman, 1983). These three causal explanations (cause-and-effect, cause, effect) are each incompatible with the truth of the conditional assertion in the original problem. Hence, they should each be judged as more probable than an explanation that rejects the categorical premise in the original problem, such as:

4. The book was sent by regular mail. (*Rejection of the categorical*)

Only the cause-and-effect explanation is a conjunction of two propositions, and so to avoid a confound in our experimental tests of the prediction, we included another conjunction in the list of explanations. It consisted of the effect paired with an antecedent that is not its cause:

5. The parcel was heavy and it went astray. (*Non-causal conjunction*)

The conjunction makes sense, but the first clause is not a cause of the effect in the second clause. Such a non-causal assertion should be rated as the least probable of the five putative explanations.

In what follows, we describe two experiments that corroborated the prediction of the order in which individuals would rank the probabilities of the five sorts of explanation.

Experiment 1:
The rank order of the probabilities of five sorts of explanation

Experiment 1 tested the rank-order prediction, and so the participants' task was to rank five sorts of explanations of inconsistencies in terms of their probabilities. In order to develop the materials, we carried out a pilot study in which we gave 10 student volunteers at the University of Padua, a series of 20 problems, i.e., four problems in each of five domains. Each problem consisted of a general conditional assertion, which was plausible, an assertion of the antecedent of this conditional, and a denial of its consequent. The participants had to give a single explanation in their own words of why each consequent had not occurred. We present here one example of a typical problem from each of the five domains (translated from the Italian):

1. The physical domain (The winter problem):
   If the air pressure is low on a winter's night then it is cloudy.
   Yesterday winter's night, the air pressure was low, but it was not cloudy. Why not?
2. Physiological domain (The aerobics problem):
   If you do aerobic exercises regularly then you strengthen your heart.
   Gino did aerobic exercises regularly, but he did not strengthen his heart. Why not?
3. Mechanical domain (The car problem):
   If the engine of a car is tuned in this special way, then its consumption of gas is reduced. The engine of this car was tuned in this special way, but its consumption of gas was not reduced. Why not?
4. Psychological domain (The amnesia problem):
   If one has a heavy blow on the head, then one forgets some past events.
   The woman had a heavy blow on the head, but she did not forget any past events. Why not?
5. Social domain (The hotel problem):
   If the hotels increase their rates for rooms, then their customers decrease.

> The hotels have increased their rates for rooms, but their customers have not decreased. Why not?

The participants generated a variety of different explanations for each of the problems with a mean of 4.75 explanations per problem. We classified the explanations, and determined which was the most frequent explanation for each problem. The results corroborated the model theory's prediction that explanations of inconsistencies of this sort should tend to rule out the conditional premise rather than the categorical premise. There may, of course, be other explanations for this phenomenon (see, e.g., Elio & Pelletier, 1997). The participants' explanations rejected the conditional premise on 90% of trials and the categorical premise on 10% of trials (Binomial $p = .5^{20}$, by materials). On four trials out of the 200 problems, the participants were unable to come up with an explanation. The participants' responses provided us with the materials for our main experiment.

Method

We tested 20 new volunteers, who were undergraduates in the Department of Psychology at the University of Padua. They had not participated in the pilot study. Each participant ranked the probabilities of a series of seven explanations for each of 20 different scenarios in which an inconsistency occurred between premises and outcome. These explanations included the five crucial assertions: cause-and-effect, cause alone, effect alone, rejection of the categorical, and a non-causal conjunction; and they also included two filler items designed to make it harder for the participants to discern the systematic pattern in the sorts of putative explanations. We derived the explanations from the pilot study, which provided us with chains of cause and effect that resolved the inconsistencies, and with rejections of categorical premises. We paired the same effects with non-causal antecedents to produce the non-causal conjunctions. The two fillers were explanations that only a few participants had spontaneously produced in the pilot study.

The seven explanations were assigned in a different random order to each of 20 problems, and the problems were printed in booklets in a random order. The 20 problems were from four scenarios in each of the five domains: physical, physiological, mechanical, psychological, and social problems. The instructions stated that the participants' task was to evaluate explanations of some surprising outcomes to brief stories. The participants were told to assign the number "1" to the explanation that seemed to be most probable, the number "2" to the next most probable explanation, and so on, until they had dealt with all seven explanations. They were also told that they could take as much time as they needed to complete the task.

Results and Discussion

The sums of the ranks of the five sorts of explanation from all the participants and all the stories yielded the following overall rank order of probabilities, where “1” equals the most probable explanation:

Cause-and-effect:	1.00
Cause:	2.45
Effect:	3.18
Rejection of the categorical:	3.38
Non-causal conjunction:	5.00

The corroboration of the predicted trend over the ranks was highly reliable (Page’s L = 1079, z = 7.98, p < 3 times in 10^7). Nine out of the 20 participants had overall ranks for the 20 problems that coincided exactly with the prediction. Given that there are 5! (= 120) possible rank orders – a conservative assumption because it does not allow for ties, the probability of 9 cases out of 20 coinciding with the prediction is highly significant (Binomial p < 3 times 10^{13}). Even in the absence of an a priori prediction, the participants showed a high correlation one with another, yielding the overall order as above (Kendall’s co-efficient of concordance W = .85, X^2 = 67.95, df = 4, p << .001).

Table 1 presents the rank orders of the explanations for each of the 20 problems. We carried out Page’s L test on the ranks for each story. The values of L ranged from 957 to 1083 with significance levels ranging from z = 2.55, p < . 01 to z = 8.18, p < < .00001. As Table 1 shows, six of the rank orders coincided precisely with the predicted rank order (Binomial p < 2 times 10^{12}). We computed Kendall’s coefficient of concordance, W, for each problem in order to determine whether the participants tended to agree amongst themselves about the relative probability of the five explanations: W ranged from .26 to .85 (with X^2 with 4 df ranging from 20.6, p < .01, to 67.6, p << .001). Hence, even with a statistical test that does not depend on an *a priori* prediction, the results corroborated the prediction.

The model theory predicts that a plausible explanation of an inconsistency should describe a cause and an effect that resolves the inconsistency. The results corroborated this prediction. The greater probability of this explanation over one that states only the cause, and one that states only the effect, are strong instances of the “conjunction fallacy” (Tversky & Kahneman, 1983), because the participants rated a conjunction as more probable than *either* of its constituents. Previous studies of the fallacy have usually shown that a conjunction is rated as more probable than *one* of its constituents (see Hertwig & Chase, 1998).

Table 1. The overall order of the means of the rankings of the probability of the five sorts of explanation for the twenty problems in Experiment 1.

For each problem, the table shows the value of Page's L, z, its one-tail probability on the standard normal distribution, and the overall rank order of the five sorts of explanation: CE (cause-and-effect), C (cause), E (effect), R (rejection of the categorical), and NC (non-causal conjunction)

	Problems	L	z	p	Overall order of mean ranks of probability
	Physical				
1.	Tectonics	1059	7.1	<< .00001	CE E C R NC
2.	Explosion	1069	7.5	<< .00001	CE C E R NC
3.	Weather	988	3.9	< .00004	CE E C NC R
4.	Melting	1062	7.2	<< .00001	CE C R E NC
	Physiological				
5.	Snake bite	1061	7.2	<< .00001	CE C R E NC
6.	Diet	957	2.6	< .006	CE R E C NC
7.	Indigestion	999	4.4	<< .00001	CE R E C NC
8.	Aerobics	973	3.3	< .0005	R CE C E NC
	Mechanical				
9.	Car	1016	5.2	<< .00001	CE R C E NC
10.	Reactor	969	3.1	< .002	CE R C R NC
11.	Pistol	1035	6.0	<< .00001	E CE C R NC
12.	Camera	1002	4.6	<< .00001	CE R C E NC
	Psychological				
13.	Forgetting	1049	6.7	<< .00001	CE C E R NC
14.	Anger	1066	7.4	<< .00001	CE C E R NC
15.	Liking	978	3.5	< .0003	CE NC E C R
16.	Anxiety	1031	5.9	<< .00001	CE E R C NC
	Social				
17.	Politics	1083	8.2	<< .00001	CE C E R NC
18.	Banks	998	4.4	<< .00001	CE C R NC E
19.	Hotels	1027	5.7	<< .00001	CE C E R NC
20.	Party	1037	6.1	<< .00001	CE C E R NC

Experiment 2:
Explanations and biconditionals

Consider again the problem of the missing book, and suppose that in place of a conditional, we used instead a biconditional ("if and only if"):

If and only if the book was sent by an express service, then it will

> arrive in two days.
> The book was sent by an express service, but it didn’t arrive in two days. Why not?

As we pointed out earlier, normal conditionals are consistent with three possibilities, but individuals tend not to represent them all explicitly. In contrast, a biconditional is consistent with only two possibilities:

express service	arrive in two days
¬ express service	¬ arrive in two days

Individuals should be more likely to represent both possibilities in explicit models (for evidence corroborating this prediction, see, e.g., Johnson-Laird & Byrne, 1991). The inconsistent fact (it didn’t arrive in two days) in the problem matches the second of these models, and so reasoners should be more likely to accept the biconditional and instead to reject the categorical premise. The same argument applies if instead of an indicative conditional, we used one that could be interpreted in a counterfactual way:

> If the book had been sent by an express service, then it would have arrived in two days.

Such counterfactuals also tend to elicit two explicit mental models, one of the counterfactual possibility and the other of the facts of the matter (Johnson-Laird & Byrne, 1991):

express service	arrive in two days [counterfactual possibility]
¬ express service	¬ arrive in two days [fact]

Byrne and her colleagues have corroborated this account experimentally (see, e.g., Byrne, 2004).

Experiment 2 was accordingly a replication of the previous study, but we compared indicative conditionals with biconditional counterfactuals – the latter designed to maximize the chances that individuals would construct two explicit mental models, and so they should tend to prefer explanations that ruled out the categorical premises. These problems were of the following form, combining a biconditional and a counterfactual:

> If and only if the book had been sent by an express service, then it would have arrived in two days. The book was sent by an express service, but it did not arrive in two days. Why not?

Method

Ten prospective students at the Scuola di Sant’Anna of Pisa carried out the rank-ordering task on two versions of each of ten of the original twenty scenarios. In one version, each scenario was the same as in Experiment 1, that is, it was based on an indicative conditional. In the other version, each scenario was based on a counterfactual biconditional. The two versions of each scenario were presented in separate blocks in a counterbalanced order, with an interval of three hours between the two blocks. We selected ten scenarios to which the previous participants had assigned a low probability to the rejection of the categorical. The ten scenarios were presented in different random orders to each participant.

Results

The results with the indicative conditionals replicated the findings of Experiment 1. But, the participants ranked the rejection of the categorical with a higher probability in the problems based on counterfactual biconditionals (an overall rank of 2.1) than in the scenarios based on indicative conditionals (an overall rank of 4.2). All ten participants fit this pattern (Binomial $p = .5^{10}$); and all ten scenarios fit this pattern too (Binomial $p = .5^{10}$). Hence, the results confirmed our conjecture.

General Discussion

Individuals create diagnoses to resolve inconsistencies. The model theory postulates that they construct a causal chain from a cause to an effect, which explains the origin of the inconsistency. The theory predicts that such a chain is more plausible than a cause alone, which in turn is more plausible than an effect alone. It also predicts that when the fact yielding the inconsistency is incompatible with the one explicit mental model of the conditional, individuals prefer an explanation that rules out the conditional in comparison with an explanation that rules out the categorical premise. Experiment 1 corroborated these predictions. When the conditional premise was changed to a biconditional with a counterfactual interpretation, individuals should be more likely to construct two explicit mental models, and one of them matches the inconsistent fact. Hence, individuals should now have an increased tendency to accept explanations that rule out the categorical premise. Experiment 2 corroborated this prediction. A subsequent unpublished experiment showed that the effect also occurred with indicative biconditionals. The preference for an explanation in the form of a cause and an effect is not merely an example of a conjunction fallacy (Tversky & Kahneman, 1983), but it also refutes a commonly held view among philosophers. Since James (1907), they have argued that individuals make minimal changes to their beliefs in order to accommodate new facts that are inconsistent with them. Our results show that this view is false: individuals prefer explanations that call for them to accept two new beliefs – a cause and an effect – that resolve inconsistencies.

References

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

Byrne, R. M. J. (2004). *The Rational Imagination: How People Create Alternatives to Reality*. Cambridge, MA: MIT.

Dieussaert, K. De Neys, W., & Schaeken, W. (2005). Suppression & belief revision, two sides of the same coin? *Psychologica Belgica, 45* (1), 1-18.

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

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

Hertwig, R. & Chase, V. M. (1998). Many reasons or just one: How response mode affects reasoning in the conjunction problem. *Thinking & Reasoning, 4,* 319-352.

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

Johnson-Laird, P. N. & Byrne, R. M. J. (1991). *Deduction*. Hillsdale, NJ: Erlbaum.

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

Legrenzi, P., Girotto, V., & Johnson-Laird, P.N. (2003). Models of consistency. *Psychological Science*, *14*, 131-137.

Revlin, R., Calvillo, D., & Ballard, S. (2005). Counterfactual reasoning: Resolving inconsistency before your eyes. *Psychologica Belgica, 45* (1), 47-56.

Tversky, A. & Kahneman, D. (1982). Causal schemas in judgements under uncertainty. In Kahneman D., Slovic, P. & Tversky, A. (Eds.), *Judgement Under Uncertainty: Heuristics & Biases* (pp. 117-128). Cambridge: Cambridge University Press.

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