Anomalous Conditional Judgments and Ramsey's Thought Experiment

John M. Miyamoto
Department of Psychology
University of Washington

James W. Lundell
Hewlett-Packard

Shihfen Tu
Department of Psychology
University of Washington

The Stalnaker/Lewis semantics for counterfactual conditionals is based on a thought experiment proposed by Frank Ramsey. We show that intuitive judgments of the truth of counterfactuals violate predictions derived from the Stalnaker/Lewis semantics. The pattern of violations suggests that the process of counterfactual reasoning follows a different pattern from the process implicit in Ramsey's thought experiment.

A counterfactual conditional is a statement of the form, "If P were true, then Q would be true", where P is a proposition that is known to be false, and Q is another proposition. For example, "If Richard Nixon had not resigned from the presidency, he would have been impeached", is a counterfactual conditional. The English philosopher, Frank Ramsey, proposed an influential heuristic analysis of conditional statements which was meant to apply to counterfactuals as well as other forms of conditionals.

> In general we can say with Mill that 'If *p*, then *q*' means that *q* is inferrible from *p*, that is, of course, from *p* together with certain facts and laws not stated but in some way indicated by the context. (Ramsey, 1931, p.248).

Stalnaker interpreted Ramsey's approach as a sequence of inferential steps applied to a knowledge base: To evaluate the truth of a counterfactual conditional,

> First, add the antecedent (hypothetically) to your stock of beliefs; second, make whatever adjustments are required to maintain consistency (without modifying the hypothetical belief in the antecedent); finally, consider whether or not the consequent is then true. (Stalnaker, 1968).

We will say that a counterfactual conditional is evaluated by a *Ramsey thought experiment* if the cognitive process by which it is evaluated proceeds through the steps described in the Stalnaker quotation. The heuristic model of the Ramsey thought experiment has been central to investigations of counterfactuals from the diverse standpoints of analytical philosophy (Goodman, 1947, 1965), intensional logic (Stalnaker, 1968, 1984; Lewis, 1973), cognitive psychology (Johnson-Laird, 1986; Rips & Marcus, 1979), and artificial intelligence (Ginsberg, 1986).

It is the thesis of this paper that the cognitive processes underlying intuitive counterfactual reasoning are quite different from the Ramsey thought experiment. We will argue that a counterfactual is evaluated by mentally constructing two alternatives, one alternative in which the antecedent and consequent are both true, and a second in which the antecedent is true and the consequent is false; a counterfactual appears to be true to the extent that the first alternative is more plausible than the second. Our arguments are based on the results of psychological experiments. We first describe a theory of counterfactuals due to Stalnaker (1968, 1984) and Lewis (1973), and derive testable relations among counterfactuals from this theory. These relations are also implied by other theories that elaborate the heuristic of the Ramsey thought experiment. We report the results of an experiment testing whether intuitive counterfactual judgments exhibit the predicted relations. To anticipate our results, intuitive counterfactual judgments violated predictions derived from the Stalnaker/Lewis analysis.

THE STALNAKER/LEWIS THEORY OF COUNTERFACTUAL SEMANTICS

The Stalnaker/Lewis theory is developed within the framework of possible worlds semantics (Kripke, 1963; Lewis, 1973; Montague, 1974). Possible worlds are abstract entities relative to which propositions have truth values. The truth values of propositions can differ from one possible world to the next. The actual state of the world is treated as one world in the set of possible worlds. Let R denote the actual world, and let α and β denote other possible worlds. The Stalnaker/Lewis theory postulates the existence of a measure of similarity, S, between the actual world and other possible worlds[1]; let $S(\alpha, R) > S(\beta, R)$ indicate that α is more similar to the actual world than β. For example, a world in which Richard Nixon did not resign and he remained under political attack is more like the actual world than a world in which he did not resign and the political attacks spontaneously ceased.

We say that α is a P-world if the proposition P is true in α. According to Stalnaker and Lewis, for any P, there exists a set τ_P that satisfies two conditions:

(a) τ_P is a set of P-worlds, i.e., P is true in every world in τ_P,

(b) Every P-world in τ_P is more similar to R than any P-world not in τ_P. In other words, $S(\alpha, R) > S(\beta, R)$ whenever $\alpha \in \tau_P$, $\beta \notin \tau_P$, and α and β are P-worlds.

If α is a world in τ_P, we will say that α is a maximally similar P-world. Stalnaker (1968, 1984) maintained that for any P, there is a unique most similar world in which P is true, i.e., τ_P always contains a single possible world. Lewis (1973) proposed that there might be several different worlds in which P is true, all of which are similar to R to the same maximum degree. We will follow Lewis in postulating that τ_P may contain more than one world.

Let $P \rightarrow Q$ denote the statement, "If P were true, Q would be true." According to Stalnaker/Lewis, $P \rightarrow Q$ is true if and only if either

C1. There is no world in which P is true, or

C2. Q is true in every world in τ_P.

Clause C1 covers the trivial case where P is logically false, e.g., "If $2 + 2 = 3$, the national debt would be eliminated" is true because $2 + 2 = 3$ is false in every world. Clause C2 captures the primary intuition behind the Stalnaker/Lewis theory. To decide whether $P \rightarrow Q$ is true, we should consider the P-worlds that are most similar to R, and ask whether Q is true in all such worlds. For example, was Richard Nixon impeached in all maximally similar worlds in which

---

1 S is assumed to be an ordinal measure of similarity defined on all pairs of worlds and not merely on pairs of the form (α, R). We will not need these more general relations here.

he did not resign? If so, we can assert that Richard Nixon would have been impeached if he had not resigned.

IMPLICATIONS OF THE STALNAKER/LEWIS THEORY

Let X AND Y denote the truth functional conjunction of X and Y, and let X OR Y denote the truth functional disjunction of X and Y. The Stalnaker/Lewis theory implies the following:

Proposition 1: For any A, X and Y, if $A \rightarrow X$ AND Y is true, then $A \rightarrow X$ and $A \rightarrow Y$ are true.

Proposition 2: For any A, X and Y, if $A \rightarrow X$ is true, then $A \rightarrow X$ OR Y is true, and if $A \rightarrow Y$ is true, then $A \rightarrow X$ OR Y is true.

Proposition 3: For any A, X and Y, if $A \rightarrow X$ AND Y is true, then A AND $X \rightarrow Y$ and A AND $Y \rightarrow X$ are true.

Propositions 1 3 are obviously true by clause C1 if A is false in all possible worlds. Therefore we will only consider cases where A is true in at least some possible worlds. To prove proposition 1, suppose that $A \rightarrow X$ AND Y is true. Then X AND Y is true in every world in τ_A; hence, X is true in every world in τ_A, and Y is true in every world in τ_A. Therefore $A \rightarrow X$ and $A \rightarrow Y$ are both true. To prove proposition 2, suppose that $A \rightarrow X$ is true. Then X is true in every world in τ_A. Hence X OR Y is true in every world in τ_A, so $A \rightarrow X$ OR Y is true. To prove proposition 3, suppose that $A \rightarrow X$ AND Y is true. Then X AND Y is true in every world in τ_A. But A is true in every world in τ_A by definition of τ_A, and X is true in every world in τ_A because X AND Y is true in every world in τ_A. Therefore A AND X is true in every world in τ_A. Thus $\tau_{A \text{ AND } X}$ must equal τ_A for if not, A AND X would be true at worlds that are more similar to R than the worlds in τ_A, contradicting that τ_A contains the maximally similar A-worlds. Hence Y is true at every world in $\tau_{A \text{ AND } X}$ (= τ_A). Hence A AND $X \rightarrow Y$ is true by clause C2 of the Stalnaker/Lewis theory.

EXPERIMENTAL TEST OF PROPOSITIONS 1, 2 AND 3

Our general procedure is to present subjects with a background story followed by a series of counterfactual statements. Subjects are asked to rate the statements for "how true or false they seem to be". The critical statements have the forms, $A \rightarrow X$, $A \rightarrow Y$, $A \rightarrow X$ AND Y, $A \rightarrow X$ OR Y, and A AND $X \rightarrow Y$. We assume that if one counterfactual is implied by a second counterfactual, the former counterfactual should receive the higher rating, for any evidence or argument that supports the latter counterfactual must also support the former. The response patterns predicted by propositions 1 - 3 are summarized in Table 1.

Miyamoto and Dibble (1986) tested propositions 1 and 2, and found violations of both propositions. The violations were analogous to conjunction and disjunction fallacies previously observed in subjective probability judgment (Morier & Borgida, 1984; Tversky & Kahneman, 1983). Propositions X and Y were chosen such that X would have been a representative outcome and Y an unrepresentative outcome relative to a background story and a counterfactual antecedent A. Statistically reliable violations of proposition 1 and 2 were found in the degree of truth ratings; $A \rightarrow X$ was rated higher than $A \rightarrow X$ OR Y, and $A \rightarrow X$ AND Y was rated higher than $A \rightarrow Y$. The present study extends these findings in three ways. First, we test proposition 3 as well as replicating the tests of propositions 1 and 2. Second, we introduce a minor procedural

TABLE 1

	Predicted Relations in Rated Truth		
Proposition 1	A → X	≥	A → X AND Y
Conjunction Test	A → Y	≥	A → X AND Y
Proposition 2	A → X	≤	A → X OR Y
Disjunction Test	A → Y	≤	A → X OR Y
Proposition 3	A AND X → Y	≥	A → X AND Y
Conditionalization Test	A AND Y → X	≥	A → X AND Y

alteration that controls against an alternative explanation to be described below. Third, and most important, whereas Miyamoto and Dibble (1986) noticed only that violations of propositions 1 and 2 contradict the Stalnaker/Lewis theory, we emphasize that violations of propositions 1 - 3 are inconsistent with the serial inferential process of the Ramsey thought experiment. Thus we are able to identify the information processing implications of these results.

EXPERIMENTAL METHOD

Subjects read a background story concerning a couple, the Conley's, and their decision whether to vacation in New York City or the Canadian Rockies. For brevity, we will omit the story, but the main points are as follows. The Conley's eventually decided to vacation in the Rockies. From their discussion of the New York option, however, it is clear that they were very interested in visiting art museums, lukewarm with respect to attending the opera, very interested in hearing live jazz, and not interested in taking walks in Central Park. Two sets of counterfactual statements were constructed to test propositions 1 3. These sets were:

Set 1

A → X: If the Conley's had vacationed in New York, they would have visited art museums.
A → Y: If the Conley's had vacationed in New York, they would have attended the opera.
A → X AND Y: If the Conley's had vacationed in New York, they would have visited art museums, and they would have attended the opera.
A → X OR Y: If the Conley's had vacationed in New York, they would have visited art museums, or they would have attended the opera, or both.
A AND X → Y: If the Conley's had vacationed in New York and visited art museums, they would also have attended the opera.

Set 2

A → X: If the Conley's had vacationed in New York, they would have heard outstanding live jazz.
A → Y: If the Conley's had vacationed in New York, they would have gone for late evening walks in Central Park.
A → X AND Y: If the Conley's had vacationed in New York, they would have heard outstanding live jazz, and gone for late evening walks in Central Park.
A → X OR Y: If the Conley's had vacationed in New York, they would have heard outstanding live jazz, or gone for late evening walks in Central Park, or both.
A AND X → Y: If the Conley's had vacationed in New York and had heard outstanding live jazz, they would have gone for late evening walks in Central Park.

TABLE 2

	Median Rating	
	Set 1	Set 2
A → X	26.0	24.0
A → Y	3.0	15.0
A → X AND Y	11.0	19.0
A → X OR Y	18.0	21.0
A AND X → Y	5.0	15.0

TABLE 3

Percentage of times the row rating exceeded the column rating;
* indicates p < .05, two-tailed sign test; ** indicates p < .01, two-tailed sign test.

Set 1	A → Y	A → X AND Y	A → X OR Y	A AND X → Y
A → X	97% **	96% **	90% **	97% **
A → Y		12% **	5% **	38%
A → X AND Y			30% **	91% **
A → X OR Y				97% **
Set 2	**A → Y**	**A → X AND Y**	**A → X OR Y**	**A AND X → Y**
A → X	90% **	88% **	72% **	92% **
A → Y		30% **	14% **	61%
A → X AND Y			20% **	80% **
A → X OR Y				90% **

Each subject read the background story and rated the statements for "how true or false they seem based on the information in the preceding story and whatever else you know about the world." The 10 statements in sets 1 and 2 were mixed with 15 additional counterfactual statements concerning related topics. Four random orderings of the statements were used in the experiment; subjects were randomly assigned to one of the four orderings. Ratings were made by placing a mark on a horizontal line that was labeled "Absolutely true" at one end, and "Absolutely false" at the other end. Intermediate positions on the line indicated intermediate degrees of truth. Responses were coded on a scale from 1 (= absolutely false) to 30 (= absolutely true).

RESULTS AND DISCUSSION

Subjects were 70 University of Washington undergraduates (mean age = 20.3, SD age = 3.16). None of the subjects had had a course in logic. There were no important differences between subjects receiving the different orderings of the statements, so results will be pooled across the orderings.

Table 2 lists the median ratings for the five statements in sets 1 and 2, and Table 3 lists the results of sign tests for differences between pairs of statements. For both sets of statements, the conjunction tests yielded violations of proposition 1. The A → Y statement received significantly lower ratings than the A → X AND Y statement ($p < .01$). For both sets of statements, the disjunction tests yielded violations of proposition 2. The A → X statement received significantly higher ratings than the A → X OR Y statement. Finally, for both sets of statements, the conditionalization tests yielded violations of proposition 3. The A → X AND Y statements received significantly higher ratings than the A AND X → Y statement ($p < .01$).

The conjunction and disjunction tests replicate the findings of Miyamoto and Dibble (1986), but also contribute a useful extension of their findings. Whereas the disjunctive statements in Miyamoto and Dibble (1986) did not explicitly indicate whether the disjunction was inclusive or exclusive, the disjunctive statements in the present study tested whether X or Y "or both" would have occurred if A had occurred. If "OR" is regarded as an exclusive disjunction, the finding that A → X is rated higher than A → X OR Y is consistent with the Stalnaker/Lewis theory. Thus, Miyamoto and Dibble (1986) was open to the criticism that the purported violations of proposition 2 were spurious because some subjects may have interpreted the disjunctive statements as exclusive disjunctions. The present study is not open to this objection. It might also be objected that subjects may interpret the A → Y statement as A → Y BUT NOT X because A → Y is perceived as contrasting with A → X AND Y. This objection has already been raised with respect to conjunction fallacies in probability judgment (Marcus & Zajonc, 1985; Pennington, 1984). We have explored this issue in additional experiments which cannot be reported here because of space limitations. To give the gist of our rejoinder, however, we examined the conjunction test in a between-subjects experiment where different subjects rated A → X, A → Y, and A → X AND Y. In such an experiment, one finds that A → X AND Y is still rated higher than A → Y. A between-subjects design eliminates the possibility that subjects contrast A → Y with A → X AND Y because different subjects rate the two statements. We have also presented subjects with statements that have the forms A → X, A → Y, A → X AND Y, and A → Y BUT NOT X. We find that the ratings of A → Y are generally higher than the ratings of A → Y BUT NOT X.

CONCLUSIONS

It is clear that intuitive counterfactual reasoning violates the conjunction, disjunction and conditionalization tests. We will refer to these violations as anomalous counterfactual judgments. What are the implications of these anomalies? First, we should recognize that the Stalnaker/Lewis theory was proposed as a normative theory of counterfactual inference, rather than as a descriptive theory of naive counterfactual judgment (Lewis, 1973; Stalnaker, 1968, 1984). Our results do not undermine the normative status of the Stalnaker/Lewis theory. Second, although we derived propositions 1 - 3 from the Stalnaker/Lewis theory, we believe they are consequences of most theories of counterfactual inference that are based on the Ramsey thought experiment. This is especially clear for propositions 1 and 2. If X AND Y is inferrible from the antecedent A and other contextually relevant beliefs, then surely X alone must be inferrible from A and these beliefs. Similarly, if X is inferrible from A and other contextually relevant beliefs, then surely X OR Y is inferrible from A and these beliefs. Thus propositions 1 and 2 are consequences of the inferential structure of Ramsey's thought experiment, rather than any peculiar feature of the Stalnaker/Lewis theory. Proposition 3 is also plausible within a Ramsey thought experiment, for if X AND Y is inferrible from A together with other contextually relevant beliefs, then X is inferrible from A together with these beliefs, and Y is inferrible from A and X and these other beliefs. Therefore Y is inferrible from A AND X together with other contextually relevant beliefs. Thus proposition 3 also follows from the inferential structure of Ramsey's thought experiment. Of course, the present argument is not rigorous, for the initial statement of the Ramsey thought experiment was heuristic rather than formal and precise.

The essential problem with the Ramsey thought experiment, as we see it, is that *a Ramsey thought experiment derives the consequences of a counterfactual antecedent without regard to the consequent of the specific counterfactual that is being evaluated.* Thus, counterfactuals that have identical antecedents undergo identical processing up to the point at which one tests whether the consequents are true in the belief structure derived from the antecedent and current belief. The existence of anomalous counterfactual judgments demonstrates that intuitive counterfactual inference does not proceed serially through the three steps of a Ramsey thought experiment.

We propose to analyze intuitive counterfactual inference along different lines from the Ramsey thought experiment. Suppose one is evaluating the truth of $A \rightarrow X$. We propose that the inference proceeds through five stages.

1. Add A and X to the current set of beliefs.
2. Construct the most plausible mental model in which A and X are both true. We assume that there is a subjective measure, $\mathcal{P}[A, X]$, that represents the subjective plausibility of this mental model.
3. Return to the initial belief state, and add A and NOT-X to these beliefs.
4. Construct the most plausible mental model in which A and NOT-X are both true. Let $\mathcal{P}[A, \text{NOT-}X]$ denote the subjective plausibility of this mental model.
5. Evaluate the relative plausibility of these two models, i.e., base the judgment of the truth of $A \rightarrow X$ on the ratio, $\mathcal{P}[A, X]/\mathcal{P}[A, \text{NOT-}X]$.

We will refer to the procedure defined by steps 1 5 as the Relative Plausibility model.

Space limitations prevent us from fully discussing how the Relative Plausibility model accounts for anomalous conditional judgments, but the general line of argument will be sketched here. To derive the anomalous conjunctive anomalies, assume that the plausibility, $\mathcal{P}[A, X]$, is a function of the similarity between a mental model in which A and X are true and a mental model of the present situation. As Tversky (1977) has shown, it is possible to increase the similarity of an instance by increasing the number of features it shares with a target. Thus, if X is a representative consequence and Y is an unrepresentative consequence of A, we should have that $\mathcal{P}[A, X \text{ AND } Y] > \mathcal{P}[A, Y]$. Furthermore, NOT-X is an unrepresentative consequence and NOT-Y is a representative consequence of A; hence, $\mathcal{P}[A, \text{NOT}(X \text{ AND } Y)] \leq \mathcal{P}[A, \text{NOT-}Y]$. Thus, $\mathcal{P}[A, X \text{ AND } Y]/\mathcal{P}[A, \text{NOT}(X \text{ AND } Y)] > \mathcal{P}[A, Y]/\mathcal{P}[A, \text{NOT-}Y]$. By step 5 of the Relative Plausibility model, $A \rightarrow X \text{ AND } Y$ should be rated as more true than $A \rightarrow Y$. The derivation of disjunctive anomalies is similar. The anomalous conditionalizations (Proposition 3) can be derived as follows. The relative plausibility of $A \rightarrow X \text{ AND } Y$ and $A \text{ AND } X \rightarrow Y$ is determined by the ratios, $\mathcal{P}[A, X \text{ AND } Y]/\mathcal{P}[A, \text{NOT}(X \text{ AND } Y)]$ and $\mathcal{P}[A \text{ AND } X, Y]/\mathcal{P}[A \text{ AND } X, \text{NOT-}Y]$. Assuming that $\mathcal{P}[A, X \text{ AND } Y] = \mathcal{P}[A \text{ AND } X, Y]$, the relative plaubility of $A \rightarrow X \text{ AND } Y$ and $A \text{ AND } X \rightarrow Y$ is determined by $1/\mathcal{P}[A, \text{NOT}(X \text{ AND } Y)]$ and $1/\mathcal{P}[A \text{ AND } X, \text{NOT-}Y]$. But $\mathcal{P}[A, \text{NOT}(X \text{ AND } Y)] <$

$\mathcal{P}$[A AND X, NOT-Y] because A and X are a plausible combination. Therefore A $\rightarrow$ X AND Y should be rated more true than A AND X $\rightarrow$ Y.

The Relative Plausibility model differs from the Ramsey thought experiment in that the content of the consequent influences the mental models that are constructed in the course of evaluating a counterfactual. The Relative Plausibility model accounts for anomalous conditional judgments by restructuring the cognitive process that is postulated to underly counterfactual inference, and by adopting aspects of Tversky's similarity model in the evaluation of the plausibility of mental models.

Acknowledgments: We would like to thank Tony Greenwald for commenting on an earlier draft of this work, and Jane Goodman and Elizabeth Loftus for useful discussions of counterfactual reasoning.

REFERENCES

Ginsberg, M. L. (1986). Counterfactuals. Artificial Intelligence, 30, 35-79.

Goodman, N. (1947). The problem of counterfactual conditionals. Journal of Philosophy, 44, 113-128. Reprinted in Goodman (1965).

Goodman, N. (1965). Fact, fiction, and forecast. Second edition by Bobbs-Merrill, Indianapolis.

Johnson-Laird, P. N. (1986). Conditionals and mental models. In E. C. Traugott, A. ter Meulen, J. S. Reilly, & C. A. Ferguson (Eds.), On conditionals. Cambridge, UK: Cambridge University Press.

Kripke, S. (1963). Semantical analysis of modal logics, I. Zeitschrift fur mathematische Logik und Grundlagen der Mathematik, 9, 67-96.

Lewis, D. K. (1973). Counterfactuals. Cambridge, MA: Harvard University Press.

Marcus, H., & Zajonc, R. (1985). The cognitive perspective in social psychology. In G. Lindzey & E. Aronson (Eds.), Handbook of social psychology (3rd ed.). Reading, MA: Addison-Wesley.

Miyamoto, J. M., & Dibble, E. (1986). Counterfactual conditionals and the conjunction fallacy. Proceedings of the Eighth Annual Conference of the Cognitive Science Society.

Montague, R. (1974). Formal philosopy: Selected papers of Richard Montague. Edited with an introduction by R. H. Thomason. New Haven: Yale University Press.

Morier, D. M., & Borgida, E. (1984). The conjunction fallacy: A task specific phenomenon? Personality and Social Psychology Bulletin, 10, 243-252.

Pennington, N. (1984). Technical note on conjunctive explanations. Center for Decision Research, Graduate School of Business, The University of Chicago.

Ramsey, F. P. (1931) The foundations of mathematics and other logical essays. London: Kegan Paul, Trench, Trubner & Co.

Rips, L. J., & Marcus, S. L. (1979). Suppositions and the analysis of conditional sentences. In M. A. Just & P. A. Carpenter (Eds.), Cognitive processes in comprehension. Hillsdale, NJ: Erlbaum, 185-220.

Stalnaker, R. C. (1968). A theory of conditionals. In N. Rescher (Ed.), Studies in logical theory. Oxford: Blackwell, ApQ Monograph No. 2. Reprinted in E. Sosa, Causation and conditionals. Oxford: Oxford University Press, 1975.

Stalnaker, R. C. (1984). Inquiry. Cambridge, MA: MIT Press.

Tversky, A. (1977). Features of similarity. Psychological Review, 84, 327-352.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. Psychological Review, 90, 293-315.