A Computational Model of Counterfactual Thinking: The Temporal Order Effect

Clare R. Walsh (cwalsh@tcd.ie)
Psychology Department, University of Dublin, Trinity College,
Dublin 2, Ireland

Ruth M.J. Byrne (rmbyrne@tcd.ie)
Psychology Department, University of Dublin, Trinity College,
Dublin 2, Ireland

Abstract

People generate counterfactual alternatives to reality when they think about how things might have happened differently, 'if only…'. There are considerable regularities in the sorts of past events that people mentally undo, for example, they tend to mentally undo the most recent event in an independent sequence. Consider a game in which two contestants will win £1000 if they both pick cards from the same color suite. The first player picks black and the second red and they lose. Most people spontaneously undo the outcome by thinking, if only the second player had picked black. We describe a computational model that simulates our theory of the mental representations and cognitive processes underlying this temporal order effect. The computer model is corroborated by tests of the novel predictions of our theory: it should be possible to reverse the temporal order effect by manipulating the way in which the winning conditions are described.

Counterfactual Thinking

When people reflect on past events, they tend to think not only about the events that actually happened but also about how those events might have happened differently. This tendency to construct imaginary alternatives to reality is called counterfactual thinking (e.g., Kahneman & Tversky, 1982). For example, if your car breaks down and you are late, you might think that you would have been on time if you had had the car serviced or if you had taken the train.

Counterfactuals have been studied in philosophy (e.g., Lewis, 1973; Stalnaker, 1968), psychology (e.g., Kahneman & Miller, 1986) and artificial intelligence (e.g., Costello & McCarthy, 1999; Ginsberg, 1986). Counterfactual thinking has been implicated in many aspects of cognition and emotion. It may play a role in formulating counterexamples in reasoning (Johnson-Laird & Byrne, 1991) and in formulating sub-goals in problem-solving (Ginsberg, 1986). Counterfactuals may allow us to learn (e.g., Roese, 1994). The sorts of counterfactuals that are useful to people may also be useful to learning algorithms in artificial intelligence systems (Costello & McCarthy, 1999). Counterfactual thinking has also been linked to a range of emotions and social judgements, including blame and regret, both in the laboratory and in real-life settings.

Yet little is known about the mental representations and cognitive processes that underlie the generation of counterfactuals. Our goal in this paper is to describe a computational model that simulates a theory of the processes underlying counterfactual thinking and some experimental results that corroborate this theory.

Psychological studies of the sorts of counterfactuals that people generate indicate considerable regularities, despite the infinite number of ways that past events could have happened differently (e.g., Kahneman & Miller, 1986). People are more likely to undo exceptional than routine events (e.g., Kahneman & Tversky, 1982), actions than inactions (e.g., Byrne & McEleney, 2000), controllable than uncontrollable events (e.g., Girotto, Legrenzi & Rizzo, 1991; McCloy & Byrne, 2000) and the first event in a causal chain (e.g., Wells, Taylor & Turtle, 1987). In this paper we will focus on one important factor that influences the mutability of an event, that is, its temporal order in relation to other events, to which we now turn.

The Temporal Order Effect

Research has shown that when a series of events are independent of each other, people tend to mutate the most recent event (Byrne, Segura, Culhane, Tasso & Berrocal, 2000; Miller & Gunasegaram, 1990; Spellman, 1997). Consider a game in which two individuals are given a shuffled deck of cards, and each one picks a card from their own deck. If the two cards they pick are of the same color (i.e., both red or both black), each individual wins £1,000. Otherwise, neither individual wins anything. John goes first and picks a red card from his deck; Michael goes next and picks a black card from his deck. Thus the outcome is that neither individual wins anything (from Byrne et al, 2000). Participants tend to undo the second event, e.g., if only Michael had picked red too, when they are asked

to imagine that one of the card selections came out differently so that the players won, and this finding has been termed the temporal order effect. In addition, the second player, Michael, is usually expected to experience more guilt and to be blamed more by John. This effect has also been demonstrated in a number of practical situations, such as in judgements of fairness in an exam context (Miller and Gunasegaram, 1990) and in ranking team performance in a league (Sherman & McConnell, 1986

One possible explanation is that causality is assigned by the relative change in probability before and after an event (Spellman, 1997) although this account cannot explain the shift in focus which arises when an explicit alternative to the first event is provided (Byrne et al, 2000). An alternative explanation is that the first event in an independent sequence may be relatively immutable because it is presupposed (Miller & Gunasegaram, 1990), acting as a background against which later events are perceived (Sherman & McConnell, 1996), and playing an important contextualising role in constructing a mental representation of a factual situation (Byrne et al., 2000). Our aim is to explain why the first event is presupposed or perceived as immutable, and to do so, we will focus on the mental representation not only of the facts, but also of the counterfactual alternatives to the facts.

People may understand the card scenario by constructing a set of mental models (Johnson-Laird & Byrne, 1991), that is, mental representations that correspond to the structure of the world, and their models may represent certain aspects of the factual situation explicitly:

John red Michael black Lose

where 'John red' represents 'John picked a red card', 'Michael black' represents 'Michael picked a black card', and 'Lose' represents the outcome (Byrne et al., 2000). They may generate their counterfactual models by mutating aspects of the factual model. The fully explicit set of models is as follows:

Factual:	John red	Michael black	Lose
Counterfactual:	John red	Michael red	Win
	John black	Michael black	Win
	John black	Michael red	Lose

where separate possibilities are represented on separate lines of the diagram, and the models may be annotated to keep track of their epistemic status (Byrne & Tasso, 1999; Johnson-Laird & Byrne, 1991). The temporal order effect indicates that people construct just a subset of the possible counterfactual models:

Factual:	John red	Michael black	Lose
Counterfactual:	John red	Michael red	Win

. . .

where the three dots represents an implicit model which may be fleshed out to be more explicit if need be. Our aim is to explain why people construct this counterfactual model more than others.

TempCounterFacts: A Computational Model of the Temporal Order Effect

We will describe a computational model, called TempCounterFacts, which simulates the primary tenets of our theory of the mental representation and cognitive processes underlying the temporal order effect in counterfactual thinking (see Walsh & Byrne, 2001, for details). The program is written in LISP and it takes as input a set of facts about the card selection game (e.g., John picked red and Michael picked black) and a description of the winning conditions (e.g., if they both pick red or both pick black they win) and it generates a counterfactual alternative about how the outcome could have turned out differently (e.g., they would have won if Michael had picked red).

We suggest that the counterfactual context, that is, the representation of the conditions leading to a counterfactual outcome, such as the conditions under which a contestant would have won or lost a card game, can in themselves provide an explicit alternative to a factual event. This possibility has not been explored systematically before. Our suggestion is that people imagine a counterfactual scenario by changing their model of the facts to fit with their model of the winning conditions. They may select the first element of the factual model, e.g., John picks red, and find a match for it in the winning models. They may consider only this model as a possible counterfactual and conclude that if Michael had picked red they would have won. We suggest that the generation of a counterfactual alternative is driven not only by the 'bottom-up' facts of the actual situation but also by 'top-down' expectations derived from the counterfactual context. The program consists of three suites of functions.

1: Representing Facts and Counterfactual Context

The program begins by constructing a representation of the facts in a FactModel:

Factual: John Red Michael Black Outcome Lose

It also takes as input a set of winning conditions and it constructs the Counterfactmodels or set of models of the winning conditions:

John Red Michael Red Outcome Win
John Black Michael Black Outcome Win

These initial models represent the conditions under which the protagonists can win. Because of working memory constraints people may represent as little information as possible (Johnson-Laird & Byrne, 1991). For example, they may not explicitly represent the conditions under which the protagonists can lose (Byrne, 1997).

The program constructs models for different conditions which may be specific events, e.g., 'John picks a red card' or not, e.g., 'both players pick a red card'. It accepts four connectives: and, or (unspecified disjunction), ori (inclusive disjunction) and ore (exclusive disjunction).

2: Matching Facts and Counterfactual Context

The program attempts to match the FactModel to the set of CounterfactModels. It selects the first fact, e.g., John picked a red card, and it attempts to find a match for this event in the CounterfactModels. If an explicit match is found then it selects this model. If not, then it looks for a model which contains the negation of the first fact, e.g., John picks not-red. In the current example, it finds a match in the first model of the CounterfactModels:

John Red Michael Red Outcome Win

Once a match is found the program checks to see if the selected model is fully explicit, that is, it contains explicit information about both card selections. If the selected model is not fully explicit then the program fleshes out the models to be explicit (for details on fleshing out models and other technicalities, see Walsh & Byrne, 2001.)

The program then compares the FactModel and the selected CounterfactModel. If they match entirely then the FactModel is a winning instance. If there is only one item that is different then it uses this model to generate a new CounterAlternative model. If there is more than one difference, then the program continutes to search through the remaining CounterfactModels to find one which is more similar to the FactModel. In this way, the program ensures that minimal changes are made (see, e.g., Byrne, 1997).

3: Generating a Counterfactual Alternative

Once a counterfactual model has been selected, the program identifies the events in this model which are different from the FactModel. In the current example, it is Michael's card. It then generates a new CounterAlternative model by taking the FactModel and replacing the FactModel events with the CounterfactModel events, i.e., Michael picked Black is replaced with Michael picked Red. It describes the newly constructed CounterAlternative by generating a counterfactual conditional of the following form:

If it had been the case that: (Replaced event)
then it would have been the case that:
(Outcome Win).

The program simulates the temporal order effect, that is, it mutates the second event, when it is given the scenario in the current example, which is typical of scenarios used in such studies. However, the program also produces a novel reversal of the temporal order effect when it is given certain descriptions, as we will now describe.

Performance of the model on novel descriptions

We tested the performance of the model on several novel scenarios, with different sorts of winning conditions and different sorts of facts (see Walsh & Byrne, 2001, for full details). For example, we gave the model descriptions of a card game in which the winning conditions required the players to pick *different* colour cards and the fully explicit set of models for the winning conditions were as follows:

John Red Michael Black Win
John Black Michael Red Win

In each case, we presented the program with the same facts, i.e., John picks a black card and Michael picks a black card and they lose, and it produces the FactModel:

Factual: John black Michael Black Outcome Lose

We varied the way in which we described the winning conditions. Given the following 'black' disjunction, to describe the winning conditions:

If one or the other but not both pick a card from a black suit, each individual wins £1,000

the program constructs the following initial models:

John Black Win
Michael Black Win

and produces the temporal order effect. When one of the CounterfactModels contains an explicit match for the first fact, as in the models of the 'Black' disjunction the program selects this model. If it is not fully explicit then the program fleshes it out, relying on the footnotes to indicate how to do so. The program compares the fleshed out model to the FactModel and if the second event is different, as it is in the example, then the program uses this event to generate a new CounterAlternative model and to produce the counterfactual conditional:

If it had been the case that: (Michael not-Black)
then it would have been the case that:
(Outcome Win).

Given instead the following 'red' disjunction, to describe the winning conditions:

If one or the other but not both pick a card from a red suit, each individual wins £1,000

The program produces the following initial models:

John Red Win
Michael Red Win

and it produces a *reversal* of the temporal order effect, that is, it constructs a counterfactual scenario that undoes the first event rather than the second event. When the CounterfactModels do not contain an explicit match for the first fact, as in the models of the 'Red' disjunction, the program selects instead a model which contains the negation of the first fact, John picks not-black (which in the binary context of the color card

game, the program recognises as red). It repeats the same process described above, however in this case it is the first event which is different in the FactModel and CounterfactModel. As a result, this event is used to generate the CounterAlternative model and the conditional:

If it had been the case that: (John Red)
then it would have been the case that:
(Outcome Win).

The winning conditions are identical in both descriptions, and the facts are identical, but the description differs. As a result, the program constructs different sorts of initial models. The program simulates the assumption of the model theory that reasoners rarely construct a fully explicit set of models and their initial set of models makes some information explicit and some implicit (see Johnson-Laird & Byrne, 1991). One novel prediction of our theory is that it should be possible to reverse the temporal order effect if the way in which the winning conditions are described ensures that people construct an initial model that does not contain an explicit match to the first event. We turn now to some experimental results that corroborate this novel prediction.

Experimental Results on Temporal Order

We constructed a scenario based on the color card scenario (from Byrne et al., 2000). In a series of experiments, the facts of the players' selections remained the same: John goes first and selects a black card, Michael goes second and the card that he selects is also black, and the outcome is that both players lose. The winning conditions were also identical in each of the experiments: the players would win if they each picked different cards. We varied the description of these winning conditions (see Walsh & Byrne, 2001, for details). The experiments test our predictions that people hold this *counterfactual context* in mind from the outset and they use it to help them construct an appropriate counterfactual model.

In one experiment, we described the winning conditions as a disjunction of red cards:

If one or the other but not both pick a card from a red suit, each individual wins £1,000

and we compared it to a control description designed to eliminate the temporal order effect (see Walsh & Byrne, 2001 for details). In the experiment, the facts were that John picked black and Michael picked black and so they both lost. The two descriptions referred to exactly the same set of winning conditions, but for the 'red' disjunction, people cannot readily match the fact about the first player, John picked black, to their explicit thoughts about how the players can win. Instead the availability of an explicit alternative to the first fact, should lead them to mutate the first fact.

We tested 148 undergraduate students from different departments in the University of Dublin, Trinity College in several large groups. They took part voluntarily and were randomly assigned to the control or 'red' disjunction condition in a between subjects design. (Five participants were eliminated because they failed to follow the instructions.) Participants completed the following counterfactual mutation task:

> Please complete the following sentence. John and Michael could each have won £1,000 if only one of them had picked a different card, for instance if...

followed by several other related tasks (for details see Walsh & Byrne, 2001).

As our theory predicts, and as our cognitive model simulates, the 'red' disjunction reversed the temporal order effect. As Table 1 shows, the results indicated that for participants who mutated a single event, more mutated the first event (40%) than the second event (24%), and this difference was reliable (binomial n = 61, z = 1.79, 1-tailed $p < .04$), whereas when the same set of winning conditions were described in the control condition, the temporal order effect was eliminated. In the control condition, as many participants mutated the first event (32%) as the second event (36%, binomial n = 32, z = .18, p = .86), as we had expected (see Walsh & Byrne, 2001, for further details).

Table 1: The percentages of mutations in Experiment 1

	Control (n = 47)	Disjunction (n = 96)
Mutations		
First only	**32**	**40**
First and then Second	15	24
Second only	**36**	**24**
Second and then First	2	1
Neither	15	11

The experiment provides the first demonstration that the typical temporal order effect can be reversed, that is, participants mutate the first event in the sequence, rather than mutating the second event. The reversal depends not on the facts or on the nature of the winning conditions but on the way the winning conditions are described. Our explanation is that this description makes some information explicitly available in the mental models that reasoners construct, and renders other information implicit in the representation. An alternative to the first player's choice was made explicitly available and as a result, the temporal order effect was reversed.

Is it possible that the results show simply that the temporal order effect does not occur when the players

must pick different cards? The original temporal order effect may be an artifact of the constraint that both players must choose the same card. However in a further experiment in this series, we ruled out this possibility (see Walsh & Byrne, 2001). We showed that the temporal order effect can be observed even when players must pick different cards. We used the same scenario as described in the experiment here except that we changed the conditionals. We described the winning conditions as a disjunction of black cards:

If one or the other but not both pick a card from a black suit, each individual wins £1,000

Participants given this 'black' disjunction exhibited the standard temporal order effect. For both the 'red' and the 'black' disjunction conditions, the facts were the same: Both players picked black. The winning conditions were also the same (the contestants would have won if they picked different colored cards). The logical form of the description was the same, in that it was an exclusive disjunction. The only difference was in the reference to the color of the suit, black or red. This small difference in wording created a large difference in mutation patterns: mutations of the first event versus mutations of the second event. Our explanation is that reasoners represent the winning conditions not in a fully explicit set of models but in an initial set of models that makes some information explicit and keeps some implicit. We have called this mental representation of the winning conditions, the counterfactual context.

General Discussion

This paper provides one of the first computational simulations of counterfactual thinking. The model simulates our theory of the mental representations and cognitive processes that underlie counterfactual thinking, in the domain of the temporal order effect. One widely held view is that the mental representation of the facts are important in the generation of counterfactual alternatives. Our model makes use not only of the representation of the facts, but also of the representation of the winning conditions, which we have called the *counterfactual context*. It constructs representations that make some information explicit and leave other information implicit.

The program simulates the robust temporal order effect. However, our theory also led to a novel prediction about the reversal of the temporal order effect. In a series of experiments, we corroborated the predictions (see Walsh & Byrne, 2001). Our experiments showed that the temporal order effect can be reversed, eliminated or observed. The experiments provide the first demonstration that the temporal order effect can be reversed and that the nature of the description of the winning conditions can influence the mutability of events.

Acknowledgements

We thank Phil Johnson-Laird, Mark Keane, Orlando Espino, David Mandel, Rachel McCloy and Alice McEleney for their advice. The research was supported by Enterprise Ireland, the Irish Council for the Humanities and Social Sciences, and Dublin University. Some of the results were presented at the International Conference on Thinking in Durham, 2000.

References

Byrne, R. M. J. (1997). Cognitive processes in counterfactual thinking about what might have been. In D. L. Medin (Ed.). *The psychology of Learning and Motivation*, Vol 37. San Diego, CA: Academic Press.

Byrne, R. M. J. & McEleney, A. (2000). Counterfactual thinking about actions and failures to act. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Byrne, R. M. J., Segura, S., Culhane, R., Tasso, A., & Berrocal, P. (2000). The temporality effect in counterfactual thinking about what might have been. *Memory and Cognition, 28,* 264-281.

Byrne, R. M. J. & Tasso, A. (1999). Deductive reasoning with factual, possible and counterfactual conditionals. *Memory and Cognition,* 27, 726-740.

Costello, T., & McCarthy, J. (1999). Useful Counterfactuals. *Electronic Transactions on the Web.* Under Submission.

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

Girotto, V., Legrenzi, P., & Rizzo, A. (1991). Event controllability in counterfactual thinking. *Acta Psychologica*, 78, 111-133.

Johnson-Laird, P. N. & Byrne, R. M. J. (1991). *Deduction*. Hove, UK: Erlbau.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, 93, 136-153.

Kahneman, D., & Tversky, A. (1982). The simulation heuristic. In D. Kahneman, P. Slovic, & A. Tversky, (Eds.), *Judgement under uncertainty: Heuristics and biases* (pp. 201-208). New York: Cambridge University Press.

Lewis, D. (1973). *Counterfactuals*. Oxford: Blackwell.

McCloy, R. & Byrne, R.M.J. (2000). Counterfactual thinking and the controllability effect. *Memory & Cognition.*

Miller, D. T. & Gunasegaram, S. (1990). Temporal order and the perceived mutability of events: Implications for blame assignment. *Journal of Personality and Social Psychology,* 59, 1111-1118.

Roese, N. J. (1994). The functional basis of counterfactual thinking. *Journal of Personality and Social Psychology*, 66, 805-818.

Sherman, S. J., & McConnell, A. R. (1996). Counterfactual Thinking in Reasoning. *Applied Cognitive Psychology*, 10, 113-124.

Spellman, B. A. (1997). Crediting causality. *Journal of Experimental Psychology: General,* 126, 323-348.

Stalnaker, R.C. (1968). A theory of conditionals. In N. Rescher (Ed.), *Studies in logical theory*. Oxford: Basil Blackwell.

Walsh, C. and Byrne, R.M.J. (2001). A computational and experimental investigation of counterfactual thinking. *Manuscript in submission.*

Wells, G. L., Taylor, B. R. & Turtle, J. W. (1987). The undoing of scenarios. *Journal of Personality and Social Psychology*, 53, 421-430.