The Journal of Mind and Behavior
Summer and Autumn 2012, Volume 33, Numbers 3 and 4
Pages 195–204
ISSN 0271–0137

Counterfactuals, Belief, and Inquiry by Thought Experiment

Jonathan Leicester

The Royal Prince Alfred Hospital

The case is presented that counterfactual thinking evolved from trial of action for inquiry into current problems. Counterfactual thinking is regulated by belief. It is activated automatically by the belief that there is a problem, and terminated by the belief that a satisfactory response is found or cannot be found. The evaluation of bad outcomes is a special case, being one among many classes of problem. The other uses of counterfactual thinking, including its extension to other applications, and its prevention of repeating the same mistake, are secondary benefits. This unified view of counterfactual thinking is seen more clearly with the original definition of counterfactual from philosophy, which allows the inclusion of future-directed conditionals.

Keywords: counterfactual, belief, inquiry, reason

This paper defines a counterfactual as an "if *p,* then *q*" or "*q* if *p*" conditional where the antecedent *p* is, or is presupposed to be, false. Allowing the falsity of the antecedent to be presupposed permits the inclusion of future-directed or forward-looking conditionals. The definition is from philosophy and philosophers have often included forward-looking conditionals as counterfactuals (for examples see Blackburn, 2008; Chisholm, 1946; McDermott, 1999). The definition is necessary for my purpose, which makes use of forward-looking conditionals. I will argue that the definition is psychologically sound. For convenience I will call this the original definition. Many psychologists use the stricter definition advocated by Kahneman and by Roese, by which a counterfactual is an "if *p,* then *q*" conditional where the antecedent *p* is known to be false. I will call this the newer definition. It restricts counterfactuals to backward-looking or past-directed conditionals, since the future is not known. I will comment further on the difficult nomenclature of conditionals in the final section of this article.

I am grateful to Raymond Russ and three anonymous reviewers for helpful comments on an earlier draft of this article. Correspondence concerning this article should be sent to Jonathan Leicester, 62 Rickard Street, Five Dock 2046, NSW, Australia. Email: jonleicester@westnet.com.au

The Congruence of Future-directed and Past-directed Counterfactuals

Daniel Kahneman (1995, p. 378) has argued that "There is no psychologically interesting difference between the counterfactual conditional 'If you had loaded one more suitcase on this cart, it would have tipped over' and the conditional warning 'If you load one more suitcase on this cart it will tip over." This is the view I wish to justify, to demonstrate that the original definition of counterfactual is psychologically sound. In this instance the two conditionals, both counterfactuals by the original definition, express essentially the same thought, derived from the same unstated background knowledge. The likely context is that they are unasked-for comments by an observer to an experienced porter, the second made as the cart is loaded, the first when it is ready to be unloaded. They are both activated by closeness or near miss, from the cue of seeing the particularly fully loaded cart. They may be idle comments, or, if the porter is a novice, they may be for instruction. The forward-looking observer may feel anxious, and the backward-looking observer may feel slight relief, but it is likely that neither speaker feels any definite emotion, they may even be amused.

Backward-looking counterfactual thoughts are often involved in attributing causes, while forward-looking counterfactuals are often involved in prediction, planning, decision-making, and actions. I will return to the question of function later, to propose that both types have the same primary purpose. At first sight the different types of counterfactuals seem to be accompanied by different emotions. Backward-looking thoughts are typically accompanied by backward-looking emotions such as regret, relief, and consolation, while forward-looking thoughts are typically accompanied by forward-looking emotions such as anxiety, fear, and hope. On closer examination, it is clearly not past and future direction that determines the emotion, it is certainty or uncertainty in the thinker's mind. It just happens that the future is uncertain and the past is often known. The football fan who thinks "I hope we will win this game, if we lose our season is over" will continue to hope until he hears the result, perhaps days after the match. A prisoner may feel regret as he thinks counterfactually of the pleasures he will forgo, knowing what his future holds. There is no necessary link between counterfactual thinking and emotion. Counterfactual thoughts that are purely hypothetical, or about matters of no personal concern, are often not accompanied by any emotions.

The Evolution of Counterfactual Thinking

My proposal begins with the conjecture that counterfactual thinking evolved from trials of action. Trial of action is the most primitive form of inquiry. It is part of conditioning behavior in simple animals, with the trying of alternatives when a previous response has had bad or neutral consequences. It is how a laboratory rat solves a maze, and thirsty cattle that find one gate to the water trough closed

move along the fence to the next gate. Humans retain this old ability to use rather automatic trials of action, as a man doing a jigsaw puzzle tries the unplaced pieces and a man with an unfamiliar set of keys fiddles and experiments to open an obstinate lock. The purpose of trial of action is to find a solution for the problem that is confronting the animal or person at the time, which for convenience I will call a current problem. Trials of action are like unspoken counterfactual thoughts: if I go to the other gate, then I may get through. If I fiddle the key out a tiny bit and try again, then the door might open. It seems a small step from trials of action to counterfactual thinking. My suggestion is that human evolution has taken this step, initially because of the better ability to solve current problems that it confers.[1]

The ability to conduct inquiry by thought experiment or mental simulation, which often involves raising and testing counterfactual alternatives, is a recent development in evolution, perhaps unique to humans (Suddendorf and Corballis, 1997). Chimpanzees are good at learning but poor at making discoveries through imagination (Povinelli, 2000). The human ability with thought experiments is presumably due to the great development of the frontal lobes in evolution from ape to human. Counterfactual thinking is a frontal lobe function, and is defective in patients with frontal lobe damage (Epstude and Roese, 2008; Knight and Grabowecky, 1995). Once evolved, this new ability greatly extended the scope and power of inquiry. Trials of action deal only with the current problem. Thought experiments can also consider the distant past, the distant future, and even the purely fantastical. Though evolved as a form of thinking that aids inquiry, counterfactual expressions are used for communication in discourse.

Counterfactual Thinking and Belief

The proposal is that in a typical instance counterfactual thinking is activated automatically when a person believes that a pertinent problem has arisen, and is terminated automatically when the inquirer believes the question is answered, the problem solved, or the best alternative found (Leicester, 2008). The stronger the belief, the more compelling its effect. These automatic processes are swift and effortless and have an unconscious component. Inquiry can also be deliberate, and can be extended deliberately. Such extensions, which override the natural response, are slow and effortful. Some people use them more than others do, depending on their intelligence and cognitive style, but it is impractical to attempt them for more than a small fraction of our thinking.

---

[1]Kahneman (1995) proposed that automatic counterfactual thinking developed from the orienting reflex seen in lower animals and infants. His example is that when the doorbell makes an unusual sound this activates the counterfactual image of its normal ring. Another example is the small problem of noticing that one's shoelace has come untied. It is part of how counterfactual thoughts are activated, a topic discussed later in this article.

Pertinent problems that initiate inquiry have many forms. These include indecision about what to do, threatening situations, planning to achieve a goal, impediments to goals and plans, bad events and bad outcomes, matters of curiosity and interest, failed expectations, disappointments, tasks and questions set by other people or advice asked for, perceptions and propositions that cause doubt or surprise, and unexpected, unexplained, or unwelcome emotional feelings.

I will comment particularly on thinking about outcomes, because this has been so thoroughly studied as a separate topic. I propose that it is a special case of the general function of thinking about problems. Bad outcomes, like other problems, automatically activate inquiry with a new search for counterfactual alternatives. Since the outcome is past, the counterfactuals will usually be backward-looking. Nevertheless, the reason the counterfactuals are activated is to seek a way to fix the current problem, which they may do by identifying a mutable cause of the bad outcome. The alternatives raised are relevant to the particular outcome, and are often not useful for the future. The inquiry sometimes succeeds: some bad outcomes can be overcome or nullified. When inquiry finds that there was a better alternative but the outcome is now past remedy, then backward-looking regret, remorse, guilt, or shame may follow, or consolation, if no better alternative is found. When a better alternative is found and the problem is one that may occur again, then the counterfactual thinking can confer future benefits, but that is not its primary purpose, it is an incidental gain.

The purpose of counterfactual thinking proposed here is consistent with the main findings from studies of responses to outcomes. These findings are that bad outcomes activate counterfactual thinking much more than good outcomes do, and that the counterfactuals raised usually focus on the most mutable events and causes in the sequence that led to the bad outcome and often suggest how the outcome might have been better.

Counterfactual thinking has other applications. It is used to attribute cause and assign responsibility. This is not what it was evolved to do, and it often yields only mutable enabling causes (Byrne, 2005). People deliberately seek counterfactuals to give consolation to themselves or to others. They do sometimes reflect on good outcomes, often when the event was one they had anticipated anxiously, or when they have the strong cue of a close shave with disaster. They sometimes think counterfactually about other people's problems and other issues that do not concern them, as well as during idle wishful thinking and reverie. People are aware of the issue of prevention, and they tend to think harder about how the outcome might have been better when they expect the situation to occur again (Markman, Gavanski, Sherman, and McMullen, 1993). I propose that these applications are secondary benefits that became available once the capacity for counterfactual thinking had evolved.

Epstude and Roese (2008), in a less general application than mine, have proposed that the recognition of a problem and the negative emotions that accompany

that recognition are the key determinants that activate counterfactual thinking. Their concern was with past-directed counterfactual evaluation of outcomes and its effect on regulation of future behavior.

The Process of Counterfactual Thinking: Its Unconscious Components and Limitations

Counterfactual thinking begins with the belief that there is a problem. This step involves processing perceptions and relevant memories and expectations, and seems to begin preconsciously. Some events automatically activate the orienting reflex, drawing attention to the problem. Often the process is more subtle, the thinker may not be fully alert to the fact he has identified the problem, which he may not express in inner speech. Problems are not always recognized promptly. The next step, if the problem requires inquiry by thought experiment, is to activate or bring to conscious mind an alternative or counterfactual antecedent for testing. This is a key step, because it is preconscious and fallible. It depends on cues and association of ideas or priming, which makes how the problem is framed important, and may explain why usual, normal and routine acts and events, and the thinker's prejudices, overvalued ideas, and strongly held prior beliefs are all so readily available. The stronger the cues, the more likely the activation (crossword puzzles depend on this for their effect). There is no control over which alternatives are activated, even during deliberate inquiry, and the inquirer may be unaware of the cues he has used. The next step is to test the counterfactual possibility that has emerged, using the Ramsey test, the inquirer hypothetically adding *p* to his stock of knowledge and evaluating his belief in the satisfactory or target outcome *q,* given *p*. When he finds a counterfactual where he believes "If *p,* then probably *q*," inquiry will stop, depending on the (closely related) strength of his belief and his judgement of the probability. In this process the emergence of belief is effortless and involuntary and its effect in terminating inquiry is automatic, though in special contexts it may be deliberately overridden. The final step, when it is applicable, is to actualize *p* to obtain *q*.

Counterfactual reasoning often fails to consider some of the most pertinent alternatives — it stops too soon. One commonly suggested reason for this is the limited capacity of working memory. I have proposed that another important reason is that the process is regulated by belief, giving speed and economy to inquiry and decision, but with some sacrifice of accuracy (Leicester, 2008). Another cause of fallibility is that even humans are simply not very good at raising counterfactual possibilities. Why does the crossword answer, known perfectly well, not emerge until the cues of some of its letters are revealed, when it pops out?

The importance of cues and the unconscious element in their use was beautifully shown in an experiment by Norman Maier (1931). Maier hung two cords

from the ceiling of his laboratory. His subjects were told to tie the cords together. The difficulty was that the two cords were too far apart to reach one while holding the other. Various objects were around the room, such as poles, clamps, extension cords, and chairs, which the subject was allowed use. The process involves a counterfactual thought experiment leading to a trial of action. Subjects found some of the solutions easily, such as tying the extension cord to the end of one of the ceiling cords. After each solution Maier told the subject "Now do it a different way." Most of the subjects failed to find the pendulum solution of tying a weight to the end of one cord and swinging it at the other cord to catch it while holding the other cord. When these subjects were well and truly out of ideas Maier gave them a cue. Apparently by accident, as he walked to the window he would touch one of the cords to set it in a slight swaying motion. Over half the subjects then found the pendulum solution within the next one minute. When they were asked how they had got the idea of the pendulum solution most said they did not know, and when pressed offered wrong reasons, including one fanciful confabulation.

Conditionals and the Psychology of Reasoning

The modern approach to the psychology of reasoning developed from work by the British psychologist Peter Wason, who showed that most subjects answer incorrectly the seemingly simple reasoning tasks he designed. This is partly because the tasks are not truly simple, but involve relatively complex logic (O'Brien, 1995). Most subjects fail to activate relevant considerations, often because of premature mistaken beliefs that the solution is found.

The dual-process theory of reasoning proposes that belief belongs to the Type 1 or System 1 process, the system that is rapid, automatic, effortless, associative, and partly preconscious. It has become appreciated that conditionals are not simply true or false, they have probabilities of being true. Belief bias refers to the tendency to endorse conclusions on the basis of believability rather than on validity. It is consistent with and perhaps predictable from the process of inquiry described above.

How people use, understand, and draw inferences from counterfactuals during discourse is less relevant to this article, though pertinent to the next section on nomenclature. It is a complex topic. People commonly make logical errors, often from understanding "if" as "if and only if" — which can lead to justifiable inferences. People sometimes use conditional expressions to give emphasis or humor to strong assertions, or courtesy to otherwise blunt instructions: these uses are unrelated to conditional reasoning.

A Note on the Classification of Conditionals

The inherent difficulty of the subject, the indefiniteness of mood in English grammar, the desire for accurate descriptive names, and a bias to choose the definition best suited to each author's own interest have led to a confused nomenclature of conditionals. A particular conditional that is counterfactual by the original definition may be semifactual, factual, or prefactual by other definitions.

English has the subjunctive mood for sentences that are definitely or possibly hypothetical and the indicative mood for factual sentences. The rule is loosely applied, with the result that many contrary-to-fact conditionals are not expressed in the subjunctive mood and some conditionals that are expressed in this mood are not counterfactual (Chisholm, 1946). This implies, I think correctly, that subjunctive conditionals and counterfactuals are not the same thing, though they are still often so treated. The hypotheticality of a conditional sentence means the degree of probability that the situations it refers to, and more especially in its antecedent, have been or will be actualized. This hypotheticality is a continuum (Comrie, 1986), and the limits of counterfactuality are unclear. With the exception of deontic conditionals expressing in general terms natural or man-made laws, conditionals are not factual: their "if" ensures some hypotheticality of the antecedent (Comrie, 1986), no matter how believable or true they are on Ramsey test, or how factual they sound and look in speech and print. In consequence, if interpreted generously, most conditionals are counterfactual by the original definition.

The main reason that Kahneman and Roese abandoned the original definition seems to have been desire for an accurate descriptive name. Kahneman and Varey (1990, p. 1102) pointed out that "By definition, counterfactual statements refer to events that did not, in fact, occur," and Roese and Olson (1995, p. 1) argued that "The term counterfactual means, literally, contrary to the facts . . . . For present purposes, we restrict our use of the term counterfactual to alternative versions of past or present outcomes, although we are aware that others have also used the term to describe future possibilities." The newer definition serves well for evaluation of outcomes and attribution of causes, and has the merit of being clear to define and apply, but it obscures the commonality of backward-looking with future-directed conditionals. It makes difficulties for studies of prediction or planning, when authors have either retained the original definition or introduced the term "prefactual" for imagined future-directed cases (Gleicher et al., 1995).

Psychologists interested in how people understand and draw inferences from conditionals during discourse often base their nomenclature on the grammatical mood of the conditional, because this affects the inferences likely to be drawn. Subjunctive conditionals have stronger counterfactual implication, and are often equated with counterfactuals, while indicative conditionals are called factual

(Byrne, 2005, pp. 33–34). Byrne adopts a suggestion originally made by Goodman (1947) and treats semifactual conditionals (even if *p*, then still *q*) separately, because, unlike other counterfactuals, semifactuals imply or concede that the actual antecedent did not cause the consequent. My comment is that the counterfactual thinking is identical in each case. Whether or not the hypothetical antecedent being tested would change the consequent emerges after the testing, from the result of the Ramsey test. I believe this justifies the original and usual emphasis on the hypotheticality of the antecedent. The important causal implication of semifactuals could be indicated by calling them concessional counterfactuals. Note that all these terms are problematical as descriptive names. Accurate and satisfactory descriptive names have proved elusive, and I think the quest for them is unnecessary.

Philosophers were initially drawn to counterfactuals by the paradox that something that has not occurred and may never occur can be true. Counterfactuals present challenges to logic and theories of truth. Early authors expected that important benefits would follow from a solution of these challenges. For example, Chisholm (1946, p. 289) contended that "The philosophical problems which this question involves are fundamental to metaphysics, epistemology, and the general philosophy of science," and Goodman (1947, p. 113) posited "A solution to the problem of counterfactuals would give us the answer to critical questions about law, confirmation, and the meaning of potentiality." It may be fair to say that neither the solution nor the benefit have been forthcoming, and to conclude that the reason philosophers separated counterfactuals from other conditionals has lost much of its force.

Furthermore, the validity of the orthodox identification of indicative and subjunctive conditionals has been questioned (DeRose, 2010; Dudman, 1994). DeRose (2010, p. 7) writes in negative tone of the "many who think they can tell what camp a conditional falls in just by quickly looking at its quasi-grammatical features." Blackburn (2008, p. 82) ends his concise entry on counterfactuals in *The Oxford Dictionary of Philosophy* with the sentence "There is a growing awareness that the classification of conditionals is an extremely tricky business, and categorizing them as counterfactual or not may be of limited use." I have some sympathy for this view, and have retreated from "counterfactual" to the simpler "conditional" several times in this article. I have chosen the definition of counterfactual that suits my interest and which I believe is psychologically sound.

Conclusions

The proposal presented is that counterfactual thinking evolved from trials of action *pari passu* with the ability for inquiry by thought experiment and the ability to consider the past and the future. The primary function of counterfactual thinking is to find solutions to current problems. Its future benefit if the same

problem recurs is a secondary gain. Counterfactual thinking is regulated by belief. It is activated automatically by belief that there is a pertinent problem, and terminated by belief that a satisfactory response is found or cannot be found. Once evolved, counterfactual thinking has extended to other applications, including consideration of good outcomes, and of problems of no direct concern to the inquirer.

The proposal is shown to be consistent with the main findings from the literatures on counterfactual thinking and on the psychology of reasoning. It involves using the original definition of counterfactual, to include forward-looking conditionals.

References

Blackburn, S. (2008). *The Oxford dictionary of philosophy* (second edition, revised). Oxford: Oxford University Press.

Byrne, R.M.J. (2005). *The rational imagination. How people create alternatives to reality*. Cambridge, Massachusetts: MIT Press.

Chisholm, R.M. (1946). The contrary-to-fact conditional. *Mind, 55*, 289–307.

Comrie, B. (1986). Conditionals: A typology. In E.C. Traugott, A. ter Meulen, J.S. Reilly, and C.A. Ferguson, (Eds.), *On conditionals* (pp. 77–99). Cambridge: Cambridge University Press.

DeRose, K. (2010). The conditionals of deliberation. *Mind, 119*, 1–42.

Dudman, V.H. (1994). Against the indicative. *Australasian Journal of Philosophy, 72,* 17–26.

Epstude, K., and Roese, N.J. (2008). The functional theory of counterfactual thinking. *Personality and Social Psychology Review, 12,* 168–192.

Gleicher, F., Boninger, D.S., Strathman, A., Armor, D., Hetts, J., and Ahn, M. (1995). With an eye toward the future: The impact of counterfactual thinking on affect, attitudes, and behavior. In N.J. Roese and J.M. Olson (Eds.), *What might have been: The social psychology of counterfactual thinking* (pp. 283–304). Mahwah, New Jersey: Lawrence Erlbaum.

Goodman, N. (1947). The problem of counterfactual conditionals. *The Journal of Philosophy, 44,* 113–128.

Kahneman, D. (1995). Varieties of counterfactual thinking. In N.J. Roese and J.M. Olson (Eds.), *What might have been: The social psychology of counterfactual thinking* (pp. 375–396). Mahwah, New Jersey: Lawrence Erlbaum.

Kahneman, D., and Varey, C.A. (1990). Propensities and counterfactuals: The loser that almost won. *Journal of Personality and Social Psychology, 59,* 1101–1110.

Knight, R.T., and Grabowecky, M. (1995). Escape from linear time: Prefrontal cortex and conscious experience. In M.S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 1357–1371). Cambridge, Massachusetts: MIT Press.

Leicester, J. (2008). The nature and purpose of belief. *Journal of Mind and Behavior, 29,* 217–238.

Maier, N.R.F. (1931). Reasoning in humans: II. The solution of a problem and its appearance in consciousness. *Journal of Comparative Psychology, 12*, 181–194.

Markman, K.D., Gavanski, I., Sherman, S.J., and McMullen, M.N. (1993). The mental simulation of better and worse possible worlds. *Journal of Experimental Social Psychology, 29,* 87–109.

McDermott, M. (1999). Counterfactuals and access points. *Mind, 108*, 291–334.

O'Brien, D.P. (1995). Finding logic in human reasoning requires looking in the right places. In S.E. Newstead and J. St B.T. Evans (Eds.), *Perspectives on thinking and reasoning: Essays in honour of Peter Wason* (pp. 189–216). Hove, United Kingdom: Lawrence Erlbaum.

Povinelli, D.J. (2000). *Folk physics for apes: The chimpanzee's theory of how the world works*. New York: Oxford University Press.

Roese, N.J., and Olson, J.M. (1995). Counterfactual thinking: A critical overview. In N.J. Roese and J.M. Olson (Eds.), *What might have been: The social psychology of counterfactual thinking* (pp. 1–56). Mahwah, New Jersey: Lawrence Erlbaum.

Suddendorf, T., and Corballis, M.C. (1997). Mental time travel and the evolution of the human mind. *Genetic Social and General Psychology Monographs, 123,*133–167.