The Role of Categories in the Generation of Counterfactuals: A Connectionist Interpretation

Robert M. French and Mark Weaver
Department of Electrical Engineering and Computer Science
University of Michigan
Ann Arbor, Michigan 48109
Tel. (313) 763-5875

Keywords: counterfactuals, norm theory, connectionism, categories

Abstract

This paper proposes that a fairly standard connectionist category mode can provide a mechanism for the generation of counterfactuals -- non-veridical versions of perceived events or objects. A distinction is made between evolved counterfactuals, which generate mental spaces (as proposed by Fauconnier), and fleeting counterfactuals, which do not. This paper explores only the latter in detail. A connection is made with the recently proposed counterfactual theory of Kahneman and Miller; specifically our model shares with theirs a fundamental rule of counterfactual production based on normality. The relationship between counterfactuals and the psychological constructs of "schema with correction" and "goodness" is examined. A computer simulation in support of our model is included.

Introduction

We believe that a picture is emerging in which counterfactuals play a significant role in human cognition; they are not a mere curiosity which may be safely ignored. Humans live in a mental world where actualities are surrounded by possibilities. In other words (as Kahneman and Miller, 1986, have proposed), when people experience an event, they may also experience plausible counterfactual alternatives, and these have a profound effect on their reaction to the actual event.

A simple example is the affective difference between missing an airplane by an hour or by a minute. The outcomes are identical, but the latter case is an order of magnitude more frustrating. The explanation for this centers on the claim that in the latter case, having made the plane is a highly available counterfactual alternative, while in the former case, it is not. Functionally speaking, this affective reaction is appropriate. It "marks" situations that, if repeated, could easily have their outcomes improved the next time.

In keeping with our claim that counterfactuals are not "special purpose" phenomena, we will also propose that their production requires no ad hoc, special purpose machinery. Instead, we suggest that a fairly standard connectionist category model suffices. Furthermore, we propose that the process by which counterfactuals are generated is closely related to the cognitive mechanisms underlying "goodness" or "schema with correction".

Counterfactual Generation vs. Counterfactual Development

The distinction between counterfactual generation and counterfactual development is not a common one, but is nonetheless crucial to our argument. We see the difference between counterfactual generation and counterfactual development as being analogous to the difference between producing an acorn and its subsequent development into an oak tree. For example, Gilles Fauconnier in *Mental Spaces* (1986), is concerned with the counterfactual world (or mental space)

to which a counterfactual proposal gives rise. (A mental space is a counterfactual world set up by statements like "If I had a milllon dollars..." or "If I'd left ten minutes earlier....") Similarly, David Lewis (1973) and, more recently, Matthew Ginsberg (1986) examine counterfactuals within the framework of standard and non-standard formal logic. In contrast, our present concern begins and ends with the production of the seed, an initial counterfactual that may (but frequently does not) develop into a counterfactual "world" or "space". We will call these counterfactuals "fleeting counterfactuals".

A concrete example may help to make this clearer. Suppose you are standing at a corner near a puddle talking to a friend. A car passes and swerves slightly to avoid splashing you. You think: "Good thing that car swerved; I would have gotten soaked otherwise" and you go on talking and completely forget the incident. In this case a fleeting counterfactual is produced but then disappears without engendering a mental space. However, this fleeting counterfactual would have served as the "seed" for the development of a mental space had you gone on to reason about the consequences of having been splashed: "I just bought these pants and the dirty water would have stained them...or would it? Anyway, I would have been cold and dirty...probably should stand back from the curb a little farther...".

The important point is the minimal nature of fleeting counterfactuals. They are the result of a low-level process, an automatic by-product of activating a category. This production is *not* under conscious control (you could never *decide* to stop having the possibility of being splashed occur to you when a passing car nears a puddle). Furthermore, the production of fleeting counterfactuals does not divert already active mental processes -- in the example, the ongoing conversation need not have been disrupted. On the other hand, the development of a counterfactual mental space does engage higher level mental processes and does so in direct proportion to the degree of development of a mental space. In the preceding example, if you had gone on to develop a space, it is likely that you also would have missed some of what your companion was saying: "Sorry, I was just thinking about what would have happened had that car hit the puddle and soaked me. What did you just say?"

We believe there is a counterfactual continuum running from fleeting counterfactuals (i.e., no mental space created at all) through the creation of a very small mental space (one or two steps of reasoning within a counterfactual world) to full blown, long-duration counterfactual spaces ("What if I had married Lynn instead of Nancy?").

Normal and Abnormal Events

Kahneman and Miller divide events into normal and abnormal ones. They define an abnormal event as "one that has highly available [counterfactual] alternatives, whether retrieved or constructed" while a normal event is one that "mainly evokes representations that resemble it". Normal events, according to them, do not evoke surprise; abnormal ones do.

The perception of the abnormal feature arises as the result of a (conscious or unconscious) comparison to its normal, expected alternative. Suppose we enter an office in which everything clearly satisfies our expectations except that the desk is upside down. The inverted desk is clearly the "abnormal" aspect of the office, and it generates a counterfactual alternative, specifically the "normal" office. Kahneman and Miller claim that the abnormal aspects of a normal event are the ones that change when we counterfactualize. We claim that the first counterfactualization that occurs "slips" the abnormal feature to one more commonly associated with the prototypical category corresponding to the event.

Representation of categories

In what follows we represent categories as clusters of "feature nodes" with mutually excitatory interconnections. In addition to interconnections, feature nodes also have connections

Node 1: pug nose
Node 2: full lips
Node 3: high forehead
Node 4: glasses
Node 5: etc.
Node 6: etc.
Node 7: shoulder-length hair
Node 8: crew cut

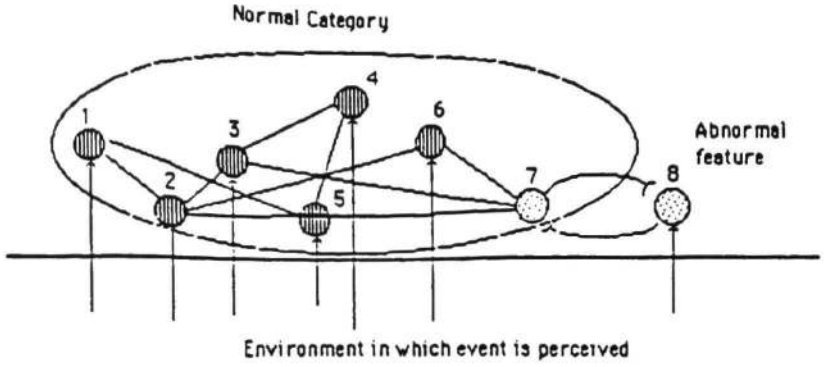

Figure 1

through which they receive environmental input. This is not unlike, for example, the schema model of Rumelhart, Smolensky, McClelland, and Hinton (1986). We represent this schematically in Figure 1 for the category "Joe".

To explain initial counterfactual production, we begin by distinguishing two different sources of activation: one external (from the environment via feature detectors) and the other internal (spreading activation from other nodes in the category). Assume that you see your friend Joe, whom you have always known to have shoulder-length hair, with a crew cut. Nodes 1 through 6 represent features that are present and sufficient to allow you to recognize your friend. Node 7 corresponds to the feature "shoulder-length hair" (always active in the past when you saw your friend) while node 8 corresponds to the feature "crew cut" (activated in the present circumstance).

Nodes 1 through 6 receive dual support, both from the environment and from other nodes within the category. Node 8 receives support *only* from the environment, while node 7 receives no external support but clearly has internal support, since your category for "Joe" has always included the feature "shoulder-length hair". Our discussion will focus on these two nodes.

Only in the situation where environmental input violates strong expectations will a strong fleeting counterfactual slip be produced. The situation is summed up in Figure 2 below.

The central role of inhibition

We will focus only on nodes 7 and 8, i.e. the long-hair and crew cut nodes respectively. We may assume that these nodes both receive strong support, the first internally, the second externally. Since people generally perceive the world as it actually is and not merely according to their expectations, we can assume an environmental bias and, consequently, that node 8 receives stronger input than node 7. (Were the opposite to be the case, misperception would occur, which does happen occasionally.) Since crew cuts and long hair are mutually exclusive features, we also assume that the corresponding feature nodes are mutually inhibitory. Initial counterfactual production (i.e., the "slip" to the normal category) would only occur once the inhibitory effect of the "crew-cut" node (8) on the "shoulder-length-hair" node (7) ceased.

The roles of habituation and short-term connection strength

There are essentially two phenomena that would cause the inhibitory effect of node 8 to cease. The most obvious is if the support from the environment simply ceased. The second, less obvious, phenomenon that would cause the activation of Node 7 to dominate that of node 8 is habituation, also called fatigue. In the first case, if environmental input ceases, our model easily demonstrates the slip to the counterfactual alternative. We decided to answer a harder question, namely: Will the slip occur even if environmental support continues unabated? Our model confirmed our intuition that the answer is yes.

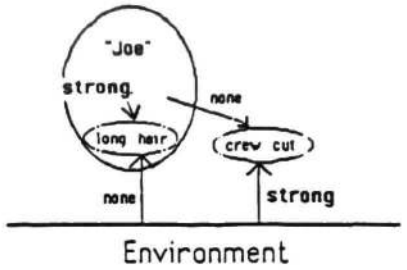

Figure 2.

The notion of neural habituation has been invoked by certain authors to explain a wide range of cognitive phenomena including decreased recall of material immediately following intensive attempts at its memorization (Pomerantz, Kaplan, and Kaplan, 1969), the Necker cube perceptual flip (Feldman,1981), flexibilty in problem solving (Levenick, 1985), etc. We claim that habituation is also an important mechanism in the production of counterfactuals.

In our model, nodes that receive on-going dual support from the category and the environment are less susceptible to the effects of fatigue than singly-supported nodes. In our example, node 8 initially receives support only from the environment. As it fatigues, its inhibiting influence on node 7 decreases. This allows the activity in the latter to rise. Node 7, increasingly active, begins to inhibit the already fatiguing node 8, further contributing to the latter's drop in activity. The activity in node 8 falls even faster, further decreasing its inhibition of node 7, and so on. Near the peak of activity in node 7 (never high enough for perception) we experience the counterfactual alternative "Joe with long hair". (The results of our simulation are consistent with this prediction.)

As we become (quite rapidly) accustomed to "Joe with a crew cut", we no longer experience the counterfactual alternative. The mechanism we propose to account for this is a short-term increase in the connection strength between the category and the new feature node. Short-term potentiation is a well-known phenomenon in neurophysiological investigations of synaptic plasticity (Goddard, 1980) and it also has been proposed as a mechanism to allow activity to briefly persist in a group of previously unassociated neural units, thus supporting the formation of new cell assemblies (Kaplan, 1970). For us, though, the importance of short-term connection strength is that it serves to temporarily strengthen the connection between the category node and the abnormal feature node.

Counterfactuals, goodness, and normality

One of the most significant results of Eleanor Rosch's extensive work with natural categories is the idea of "goodness" (Rosch, 1977). She showed that categories had no definite boundaries and that not all members are equal. Instances closer to the prototype (or most normal instance) are judged as being better than those which are most distant.

We claim that our category model provides a simple, direct coding of goodness by the overall level of activation of the nodes making up the category. Whenever a normal feature is replaced by an abnormal alternative, the normal node acts as an activity sink; it receives activation from other feature nodes but does not generate any in return. Thus the overall level of activity of the category is reduced and the instance is experienced as poorer. When the violation of expectations is strong enough, counterfactualization also occurs.

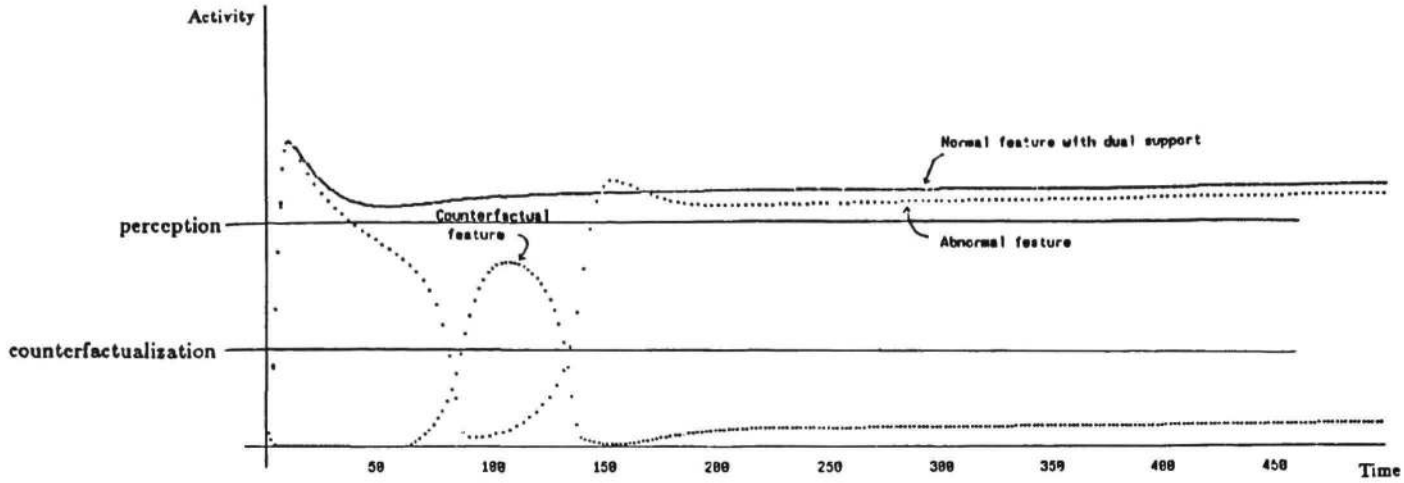

Figure 3 Activity trace showing production of a fleeting counterfactual

Counterfactualization and schema with correction

A time-honored construct in psychology is schema with correction (Woodworth, 1958). The usual context is one of asking people to redraw nonsense figures from memory. Where possible, people remember such figures as familiar figures plus an abnormal feature (eg., a square with a "funny" corner). But this is actually a case of counterfactualization. The "square" with a "funny corner" is, in fact, *not* a square at all. It *would* be a square only if the funny corner were normal. A prerequisite for schema with correction is counterfactualization and identification of an abnormality.

Simulation of the production of a fleeting counterfactual

The feature nodes in our category model do not represent single neural-level elements but rather interconnected groups or assemblies. So, a standard weighted-sum activity function was inappropriate for the simulation. Accordingly, we based our simulation on on a slightly modified mathematical model developed by Stephen Kaplan in 1970. His model was originally designed to describe the time course of activity within a single cell assembly. It allows persistent node activity (due to internal positive feedback) and, significantly, includes fatigue and short term potentiation mechanisms. We created a network of three nodes each of which would behave, in isolation, like Kaplan's cell assembly. We created links between the two nodes in competition. Initially, we did not use changing short-term connection strength between the category and the competing nodes to influence the spread of activity to these nodes. This resulted in a continual alternation of activity peaks between the normal and counterfactual nodes. When we realized that activity spreading *between nodes* was no different than *within* a node, we added a short-term connection strength mechanism to our links. Thereafter the model behaved according to our predictions (Figure 3).

Conclusion

In the world of human cognition, actualities are surrounded by a context of counterfactual possibilities. This context has a powerful effect on people's reactions to objects and events. We have chosen to study what seem to be to us the simplest, most easily isolated counterfactuals which we have called *fleeting counterfactuals*. We have shown that a slightly extended connectionist category model will produce this counterfactual behavior and thus provides the beginnings of an explanatory mechanism for the phenomenon of counterfactuals.

Appendix I

Mathematical details of the simulation

The following equations, originally developed by Stephen Kaplan in 1970, model the time course of activity in a neural net. Only very minimal modifications to the original equations were required to adapt them to our own model. (Note: F_t denotes F(t), S_t denotes S(t), etc.)

Sensitivity of connections:

This equation says, in essence, that as fatigue builds, the sensitivity (i.e., the ability to pass activation) of the affected connection drops rapidly. Sensitivity is described by:

$$\alpha_t = \frac{(L_t + S_t)(1 - F_t)^2}{K_\alpha}$$

where:

L_t and S_t are, respectively, long-term and short-term connection strength;
F_t is fatigue;
K_α is a constant designed to ensure that α_t does not exceed 1.

Change in short-term connection strength over time:

$$\Delta S_t = K_{sscale} P_t (1 - S_t)^2 - K_{sdecay} S_t$$

where:

S_t is short-term connection strength;
P_t is the activity of the net;
K_{sscale} and K_{sdecay} are constants.

Change in fatigue over time:

$$\Delta F_t = K_{fscale} P_t (1 - F_t)^2 - K_{fdecay} F_t$$

where:

F_t is fatigue;
P_t is the activity of the net;
K_{fscale} and K_{fdecay} are constants.

Input to the system:

I_t: Input to the nodes under examination was supplied by a square input function. Environmental input to node 8 was weighted somewhat higher (10%) than the category input to node 7, our reason being that environmental support is given more weight than equivalent internal support, in order to improve the probability that misperception of the environment does not occur.

Change in activity over time:

$$\Delta P_t = [P_t + (1 - P_t)I_t](1 - P_t)\alpha_t - [P_t^5 + P_t(1 - P_t)^{10}](1 - \alpha_t)$$

where:

$[P_t + (1 - P_t)I_t](1 - P_t)\alpha_t$ is the rise component and
$[P_t^5 + P_t(1 - P_t)^{10}](1 - \alpha_t)$ is the fall component of activity.

For a more detailed explanation of the derivations of these equations, see (Kaplan, 1970).

Bibliography

Fauconnier, G. (1985). Mental Spaces. Cambridge, MA: MIT Press.

Feldman, J.A. (1981). A connectionist model of visual memory. In G.E. Hinton and J.A. Anderson (Eds.), Parallel models of associative memory Hillsdale, NJ:Lawrence Erlbaum Associates.

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

Goddard, G. (1980). Components of the memory machine revisited: Hebb revisited. In P.W. Jusczyk and R.M. Klein (Eds.), The nature of thought: Essays in honor of D.O. Hebb. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kahneman, D. and Miller, D.T. (1986). Norm theory: Comparing reality to its alternatives. Psychological Review, 93 (2), 136-153.

Kaplan, S. (1970). The time course of activity in a neural net. Unpublished paper. The University of Michigan.

Levenick, J. (1986). Knowledge representation and intelligent systems. Unpublished doctoral dissertation, University of Michigan, Computer Science Department.

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

Pomerantz, J.R., Kaplan, S. and Kaplan, R. (1969). Satiation effects in the perception of single letters. Perception and Psychophysics, 6, 129-132.

Rosch, E. (1977). Principles of categorization. In E. Rosch and B. Lloyd (Eds.), Cognition and Categorization. Hillsdale, NJ: Lawrence Erlbaum Associates.

Rumelhart, D.E., Smolensky, P., McClelland, J.L., and Hinton, G.E. (1986). Schemata and sequential thought. In J.L. McClelland and D.E. Rumelhart, Parallel Distributed Processing. Cambridge, MA: MIT Press.

Woodworth, R. (1958). Dynamics of Behavior. New York: Henry Holt and Co.