ELECTRONIC NEWSLETTER ON
REASONING ABOUT ACTIONS AND CHANGE

Issue 00004 Editor: Erik Sandewall 27.6.2000

Today

Today we present questions and comments by Judea Pearl for the the ETAI submitted article by Tom Costello and John McCarthy, *Useful Counterfactuals*, as well as the authors' answers.

ETAI Publications

Discussion about received articles

Useful Counterfactuals

by **Tom Costello and John McCarthy**

Comment from: **Judea Pearl**

The title of this paper, "useful counterfactuals", seems to suggest (1) that plain ordinary counterfactuals are useless, (2) that the paper will teach us how to discriminate useful from useless counterfactuals and (3) that the paper will teach us how useful counterfactuals can be interpreted and put into use. I will start by arguing that **all** (true) counterfactuals are useful. This implies that (2) is not needed. Finally, I will discuss whether the paper is specific enough to accomplish (3).

The word "counterfactual" connotes contradiction and/or metaphysical speculation but, in fact, counterfactuals are neither contradictory nor metaphysical. Counterfactuals carry as clear an empirical message as any scientific laws, and are fundamental to them. The essence of any scientific law lies in the claim that certain relationships among variables remain invariant when the values of those variables change relative to our immediate observations. Counterfactuals, likewise, tell us what remains invariant when the world undergoes change, so, they are not different than any scientific knowledge. Every scientific law can be expressed counterfactually; for example, Ohm's law can be stated: "had the current in the resistor been I instead of I', the voltage would have been V' (=I'R), instead of V."

Thus, to say that counterfactuals are useful amounts to saying that scientific knowledge is useful. (In my IJCAI-99 paper I discussed how even a hindsighted sentence such as "had I bet differently, I would have won a dollar", which refers to a non-repeatable circumstance, conveys useful knowledge. (Paper available on www.cs.ucla.edu/ judea/))

Costello and McCarthy list several usages of counterfactuals, including the conveyance of facts, learning, and prediction, yet when we examine those usages, we find that none is distinct to counterfactuals, and each may equally be served by what we normally call "domain knowledge". Indeed, none of the sentences on pages 2-3 would turn false if we replace the word "counterfactual" with the word "knowledge". The distinct puzzling questions about counterfactuals are (1) why humans resort to this subjunctive mode of expression in conveying simple chunks of knowledge, and (2) how we should interpret counterfactual sentences so as to extract the knowledge they convey. The paper does not touch on question (1) and the answer it provides to question (2) is not presented in a form that is transportable beyond the example in which it is embedded.

The paper begins by emphasizing two notions,

1. Counterfactuals are evaluated with the aid of "approximate theories", and
2. The truth of counterfactuals depends on representation of situations in a "Cartesian space".

By "approximate theories" the authors mean: "not complete theories of the world". I agree with the ubiquity of such theories, but I fail to see what makes approximate theories particularly akin to counterfactuals. It seems to me that EVERY theory is approximate, or else we would call it a "model" i.e., a mathematical object that assigns truth value to every sentence of interest. Moreover, I fail to see what is wrong with the traditional approach of first defining the truth of a sentence in a model (i.e., "complete theory") and then, if we do not have a complete specification of a model, we say that the sentence is true in a theory just in case it is true in every model of the theory. Do the authors suggest that this approach should be abandoned when it comes to counterfactuals? If so, why? And if not, I would have liked the authors to tell us what they consider to be an appropriate MODEL for counterfactual sentences. In other words, what kind of things we must specify before we can assign truth value to a counterfactual sentence $q > p$, where p and q are arbitrary propositions. I have not found the answer in this paper, and this made the reading very difficult.

The second notion emphasized in the paper is "Cartesian space", which is a space of points defined by coordinates, such that we can always change one coordinate while keeping the others unchanged. When we can do that to every point (X, Y), it makes sense to say, "if X were 3, we would be closer to the origin", because it is possible then to infer the final location from the initial one. This Cartesian metaphor corresponds to what philosophers called "ceteris paribus" (keeping everything else constant), first suggested by J S Mills (1843). as a key element in understanding counterfactuals. Given this basic intuition, I concur with the authors that some form of ceteris paribus must govern the interpretation of counterfactuals no matter what formalism we use. However, this is only the first step. The remaining steps are to decide WHERE the Cartesian space is to be found in any given story, how to represent the points in that abstract space, how to compute the coordinate change dictated by a given counterfactual $q > p$ and, finally, how to compute the ramifications of this coordinate change on other propositions. I felt that the paper leaves these questions either unanswered or implicit in the formulation of the skiing example. If the latter is the case, I suggest that the authors cast the answers in the form of generic principles, to permit their transportation across domains.

I was unable to understand the skiing example; it is too involved and too skiing-domain-sensitive for me to master. I strongly recommend that the authors choose another example, one that leaves no ambiguities as to what theory resides in the mind of each speaker, and what the right answer is in each theory. In my IJCAI-99 paper I chose, for example, a firing squad scenario, where the entire theory can be communicated unequivocally to skiers and non-skiers alike, and where the truth value of every counterfactual sentence is obvious to all readers (e.g., that if rifleman-1 had not shot, the prisoner would still be dead). I would be glad to comment on the authors' proposed axiomatization, once it is cast in a more familiar domain, and if the authors demonstrate explicitly how counterfactual sentences can be evaluated IN GENERAL. Examples of explicit demonstrations can be found in Charles Ortiz's AIJ-9 paper, using a domino-tiles example, and in my IJCAI-99 paper (using the firing squad scenario).

I can comment on section 8, entitled Bayesian Network, with which I am somewhat more familiar. First, the title may be confusing; Bayesian networks cannot support counterfactual reasoning, for reasons described in Balke and Pearl 1994 a b (see also Causality 2000, p.33-37) The authors probably meant to discuss probabilistic Structural Equations Models (SEM), of which Bayesian networks are an abstraction. An SEM is defined as a set of deterministic functions, while a Bayesian Network is defined as a set of conditional probability constraints.

Costello and McCarthy identify two differences and one commonality between their formulation and SEM. I will explain why I find these two differences to be illusionary and the commonality to be tangential. I will then discuss what I consider to be the essential difference between the two formulations.

Illusionary Difference 1.

Costello and McCarthy write:

> *One major difference between our approach and structural equation models or Bayesian networks is that we consider arbitrary proposition, and consider these relative to a background approximate theory ....*

The structural equation approach also considers arbitrary propositions relative to a background approximate theory. Contrary to Costello and McCarthy interpretation, the structural equation approach is NOT committed to equational specification. True, the equations (or functions) are used in defining a complete causal model, but "arbitrary propositions" can be used to specify "approximate causal theories", much the same as a collection of clauses defines a "theory" in propositional calculus; while a truth assignment to all the elementary propositions defines a "model". In fact, the most common causal theories used in SEM are in the form of graphs, which make no commitment whatsoever to the functional form of the equations;

graphs merely restrict the set of arguments in each equation. Arbitrary domain constraints are not excluded from SEM, they are merely interpreted as constraints over the functional relationship between A and B.

To illustrate, suppose our approximate theory contains just one (causal) rule: "If A then B". Suppose further A and B are true. Question: Is the counterfactual "not-A > not-B" true? Answer: we dont know. We need to complete the rule "If A then B" into a function, to specify what happens to B when A is false. This completion, whether done explicitly, or my minimization, or by some other principle, turns our theory into a collections of functions, a collection which I called a "structural causal model" .

Illusionary Difference 2.

The paper states: "The other major difference is that Bayesian networks focus on the probability distribution of certain variables, rather than on facts in general." I would like to believe that the authors did not mean it literally, because this kind of psuedo-differences were used by some AI-ers in the 1980's as an excuse for not reading the probabilistic literature. Those who venture to read that literature would discover quickly that probabilistic reasoning (in SEM) proceeds in two steps: First, reasoning about "facts in general" in a deterministic theory, and second, computing the probability of those "facts in general" when we have additional knowledge on how likely the background (or "frame") facts are. In my IJCAI-99 paper, for example, I first demonstrate how to compute the truth value of the deterministic counterfactual "if rifleman-1 had not shot, the prisoner would still be dead", and ONLY THEN I go to computing the probability of this sentence, assuming that rifleman-1 is somewhat likely pull the trigger out of nervousness. The same sequence is followed in Balke and Pearl (UAI-1995) in Galles and Pearl (1997, 1998) and in my new book Causality, chapter 7 (partly on www.cs.ucla.edu/ judea/)

Thus, probabilities are options, not barriers to students of counterfactuals.

Tangential commonality

Costello and McCarthy write that their approach "can be seen to be similar to modeling systems with structured equations.. or Bayesian networks..." In their Theorem 4, they prove that a counterfactual sentence "is true in a causal model M if and only if [it] is true in the Cartesian frame MF..." What Theorem 4 states is that the evaluation of counterfactual sentences in a causal model M (according to the SEM formalism) involves a Cartesian-product-like assumption, and that one can identify the Cartesian space with the set of functions F. This is correct: structural models indeed assume ceteris paribus relative to the set of equations – when we change one equation, the others remain intact.

The reason I consider this commonality to be tangential is that I (and most people I know) take for granted the invocation of ceteris paribus in counterfactual analysis. The interesting question, in my opinion, is not whether a ceteris paribus assumption is present in a given theory of counterfactuals – such presence is inevitable – but rather, what space should we apply ceteris paribus to and how. Balke, Galles and Pearl make a specific commitment in this regard. They claim that ceteris paribus should be applied to the space of MECHANISMS (read: functions), and NOT to the space of propositions and not to the space of variables, and not to some other space that one can dream up.

Thus, if Costello and McCarthy buy this commitment, they can safely claim that their approach "can be seen to be similar to modeling systems with structured equations.. ". But the mere existence of ceteris paribus (or Cartesian product space) someplace in a system does not make their approach similar to that system.

And this brings me to the main point of my comments: Have Costello and McCarthy made this commitment (i.e., to identify the Cartesian space with a set of mechanisms)? Incidently, is anyone on this Newsletter prepared to make this commitment?

A word of caution to those who answer YES or WHY NOT: the commitment to mechanisms does not come cheap. First, it requires that we proclaim certain sentences as "mechanisms", and that we assign to those sentences a different status and a different syntactic representation than that assigned to other sentences (e.g., facts, observations, assumptions, implications) It also requires a one-to-one correspondence between mechanisms and variables (see my IJCAI-99 paper, Sections 4.4-4.5, on www.cs.ucla.edu/ judea/) I have not seen these two elements in the authors' analysis of the skiing example, but I may have overlooked them, given my ignorance of skiing instructions.

The reason I emphasize this commitment to mechanisms is that I do not believe counterfactual reasoning (or causal reasoning in general) is feasible without it. The puzzle with a counterfactual sentence, say $q > p$, is that it involves a relationship between two PROPOSITIONS, q and p, not between an action and a proposition, and yet we treat q as an ACTION. How? In order to change the actual world to satisfy q we need to translate q into some action and to decide which mechanisms are to be altered by that action. Every theory of counterfactuals ought to explicate how these decisions are made in the representational scheme employed. It is quite possible that Costello and McCarthy's theory embeds these decisions implicitly in their analysis of the skiing example (which I missed). If they did, I believe they should make them formal, general and explicit.

Here is an example of some general principles that people have proposed for counterfactual analysis. Balke, Galles and Pearl (ijcai-99) identified 3 necessary steps in the evaluation of a counterfactual sentence: 1. Abduction, 2. Action and 3. Prediction. For example, to evaluate the sentence: "if rifleman-1 had not shot, the prisoner would still be dead. we must execute the following three steps:

1. abduction: use the fact that the prisoner IS dead to infer that the captain gave the order to shoot.
2. action: alter the mechanism which originally made rifleman-1 obedient to the Captain's order.
3. prediction: test whether the prisoner is alive in the new theory, created by steps 1 and 2.

If Costello and McCarty consider these steps NOT necessary for the evaluation of counterfactuals, I would invite them to posit alternative generic steps which they do consider necessary. (In this case, I would also challenge them to evaluate the sentence above with their alternative steps – we must be concrete.) If they do consider these steps to be necessary, then I would ask them to identify where in the Costello-McCarthy formalization we find traces of these steps, and how we should go about deciding (in step (2)), what changes to make to the theory so as to accommodate the counterfactual antecedant "had rifleman-1 not shot".

Have I left out another possibility? Yes, that Costello and McCarthy consider these three steps to be necessary but not sufficient. In this case, the readers of this paper would want to learn what additional principles they deem necessary, and in what kind of theories the need for the new principles will become urgent.

The difficulty I had in reading this paper stemmed from not knowing where the authors stand on these possibilities and, consequently, I could not see the principles that the paper advocates for the analysis of counterfactuals .

I hope the authors will provide this information in a revised version.

Judea Pearl

PS. Transcript and slides of my IJCAI-99 lecture are now available on [1][`http://www.cs.ucla.edu/ sunshine/pres/ijcai99.htm`]

Comment from: **author**

1. Judea Pearl considers all counterfactuals as useful and therefore has no use for our singling out useful counterfactuals.

We thought our distinction was clear, but we have added some material to the article to make it more clear.

An example of a not demonstably useful counterfactual is "If Caesar was in charge in Korea, he would have used catapults". It is difficult to see how the truth of this might teach us something. Pearl's example of the firing squad is probably (at least 0.8) useless in most Americans daily lives. No use of this counterfactual was offered.

We wrote that a counterfactual is useful if believing it can affect behavior. Our example was "If another car had come over the hill when you passed, there would have been a head-on collision." If the driver believes it, he will be more conservative about passing in the future. While some theory is involved in accepting the counterfactual, it is basically about a single experience and an associated almost experience (i.e. the collision).

Contrast this counterfactual with Pearl's examples about the firing squad, e.g. "If A had not fired the prisoner would have died anyway." Pearl offers no suggestion about how believing it would help design better firing squads or would help the victim escape. Indeed the counterfactual is entirely derived from theory - no specific experience plays any role.

[1]Ref: http://www.cs.ucla.edu/ sunshine/pres/ijcai99.htm

Useful counterfactuals like the car example have another property. They have non-counterfactual consequences, e.g. "Passing under the conditions of the example is unsafe." This is in contrast with David Lewis's theories, where counterfactuals have no non-tautologcal non-counterfactual consequences.

We have modified the article to emphasize this aspect of useful counterfactuals.

2. Pearl considers our skiing example exotic. We admit to more experience with skiing than with firing squads. We doubt Pearl's experience is otherwise. More to the point, the counterfactual sentences about skiing have many useful non-counterfactual consequences.

3. Very likely, we should have compared our useful common sense counterfactuals with those involved in stating scientific laws. However, so far as we know, the literature relating them does not describe drawing non-counterfactual conclusions from the counterfactuals themselves. It concerns the interpretation of counterfactuals rather than their use. Maybe the defenders of historical counterfactuals discuss non counterfactual consequences.

4. Pearl asks why humans resort to counterfactuals. We think it is because of their useful non-counterfactual consequences.

5. Pearl suggests that the two differences we point out between Structured Equational Models and Cartesian Counterfactuals are illusionary.

We agree that they are not major technical differences, but are mainly differences in subject matter, and presentation. Cartesian Counterfactuals are applied to commonsense theories expressed in First Order Logic, in contrast to SEMs which are applied to other domains, and are expressed in equations.

Pearl also suggests that the theorem we give is tangential. We agree that it does not address the underlying question of where Cartesian Frames or Causal Mechanisms come from.

Tom Costello and John McCarthy

This newsletter is issued whenever there is new news, and is sent by automatic E-mail and without charge to a list of subscribers, in plaintext format. The present postscript issue is a bona fide rendering of the same information in alternative typography. Instruction for contributors and other additional information is found at the following URL:

`http://www.etaij.org/rac/`