Counterfactuals in Economics: A Commentary

Nancy Cartwright

LSE and UCSD

0. Introduction

Counterfactuals are a hot topic in economics today, at least among economists concerned with methodology. I shall argue that on the whole this is a mistake. Usually the counterfactuals on offer are proposed as *causal surrogates*. But at best they provide a "sometimes" way for finding out about causal relations, not a stand-in for them. I say a "sometimes way" because they do so only in very special -- and rare -- kinds of systems. Otherwise they are irrelevant to establishing facts about causation. On the other hand, viewed just as straight counterfactuals, they are a washout as well. For they are rarely an answer to any genuine "What if…?" questions, questions of the kind we pose in planning and evaluation. For these two reasons I call the counterfactuals of recent interest in economics, *impostor counterfactuals*.

I will focus on Chicago economist James Heckman, since his views are becoming increasingly influential. Heckman is well known for his work on the evaluation of programs for helping workers more effectively enter and function in the labor market. I shall also discuss economist Stephen LeRoy, who has been arguing for a similar view for a long time, but who does not use the term "counterfactual" to describe it. I shall also discuss recent work of Judea Pearl, well known for his work on Bayesian nets and causality, and economist/methodologist Kevin Hoover,

as well as Daniel Hausman. I shall begin with a discussion of some counterfactuals and their uses that I count as genuine, to serve as a contrast with the impostors.

Before that I need one technical remark. I shall talk about causal models. As I shall use the term, a causal model for a given system or kind of system (such as a toaster of a given make or the U.K. economy in 2003) is a set of equations that represent a (probably proper) subset of the causal principles by which the system operates. The equations are supposed to be functionally true. In addition, the quantities on the right-hand side are supposed to represent a complete and minimal set of causes for the quantity represented on the left; to signal this I use not an ordinary equal sign but rather "$c=$".

The equations may represent deterministic principles or they may contain random variables that do not represent real quantities but serve to allow for a purely probabilistic relation between a full set of causes and their effect. In this case the causal model must also specify a joint probability distribution (that I shall designate by μ) over these 'dummy' variables. For simplicity of presentation I will assume that the contributions of the different causes are additive. I also assume that causality is asymmetric, irreflexive and functionally transitive.[1] So a causal model will look like this:

(CM) $x_1 c= z_1$

$\vdots$

$x_i c= \sum_{j<i} a_{ij} x_j + z_j$

$\vdots$

$$x_n \, c = \sum_{j<n} a_{nj} x_j + z_n$$

$$\mu(z_1, \ldots, z_n).$$

The *x's* represent known quantities. The *z's* are random variables, which may either represent the net effect of unknown causes or may be dummy variables that allow for the representation of probabilistic causality. (Note that my characterization is not the same as Judea Pearl's because Pearl does not allow for purely probabilistic causation.)

1. Genuine counterfactuals

1a. The need for a causal model

Daniel Hausman tells us "Counterfactual reasoning should permit one to work out the implications of counterfactual suppositions, so as to be prepared in case what one supposes actually happens."[2] My arguments here will echo Hausman. The counterfactuals that do this for us provide genuine answers to genuine "What if…?" questions; and they play a central role throughout economics. When we consider whether to implement a new policy or try to evaluate whether a trial program has been successful, we consider a variety of literally intended counterfactual questions: "What if the policy were put in place?" "What if the program had not existed?"

These are just the kinds of questions Heckman considers in his applied work, where he is at pains to point out that the question itself must be carefully formulated. We may for instance want to

know what the wages of workers in the population at large would have been had the program not existed; more commonly we end up asking what the wages of workers *in the program* would have been. Or we may want to know what the GDP would have been without the program. We also need to take care about the contrast class: do we want to know the difference between the results of the program and those that would have occurred had no alternatives been present or the difference compared to other programs, real or envisaged?

To evaluate counterfactuals of this kind we need a causal model; and the causal model must contain all the information relevant to the consequent about all the changes presumed in the antecedent. There is no other reasonable method on offer to assess counterfactuals. We may not always produce a model explicitly, but for any grounded evaluation there must be a causal model implicit; and our degree of certainty about our counterfactual judgments can be no higher than our degree of certainty that our causal model is correct.[3]

Aside. David Lewis and his followers suppose that we need a model containing the principles by which the system operates (a *nomological* model) to assess counterfactuals but not a *causal* model. I do not agree. But it is not this distinction between a Lewis-style merely nomological model and a causal model that I want to discuss here. Rather I want to focus on the difference between the causal models that support the counterfactuals we use directly in policy deliberations and those associated with impostor counterfactuals. *End of aside.*

For purposes of evaluating a counterfactual, besides our causal model we will need to know what changes are envisaged -- usually these are changes under our control. Before that we will need to

know what changes are possible. This will depend on the structure of the system and the principles under which it operates. For this question, a causal model as I have characterized it is insufficient, for the causal model does not yet carry information about what can and what cannot be changed. I will turn to this question first, in section *1b*., then in *1c*. take up the relation between counterfactuals and the changes they presuppose.

1b. What can be changed?

Some people take there to be a universal answer to the question of what can (and should) be changed in assessing counterfactuals: every separate causal principle can be changed, leaving everything else the same, including all other causal principles, all initial values and all conditional probability distributions of a certain sort. Judea Pearl claims this; so do James Woodward and Daniel Hausman.

Hausman and Woodward defend this view by maintaining that the equations of a causal model would not represent *causal* principles if this were not true of them. I have, however, characterized the equations in such a way as to give a different job to them: they are to be functionally correct and to provide a minimal full set of causes on the right-hand side for the quantity represented on the left. The two jobs are different and it would be surprising if they could both be done in one fell swoop as Hausman and Woodward claim.

Hausman and Woodward object that the jobs cannot be different since the following is true by virtue of the very idea of causation: if a functional relationship between a set of factors

(represented by, say, $\{x_j\}$) and a different quantity (say x_e) is functionally correct and the set $\{x_j\}$ is a minimal full set of causes then it must be possible to change this functional relationship, and indeed to stop every one of the x_j from being a cause of x_e, without changing anything else. The x_j would not be causes of x_e were this not true.

I think this claim is mistaken. There is any number of systems whose principles cannot be changed one at a time without either destroying the system or changing it into a system of a different kind. Besides, this assumption does not connect well with other features of causality, described in other accounts, such as probabilistic theories, causal process theories or manipulation accounts.[4]

Pearl has another argument. He says that this assumption is correct because otherwise counterfactuals would be ambiguous. As far as I can tell, the argument must go like this:

i. Before we can evaluate $c \,\square\!\!\rightarrow e$ we must know how *c* will change, otherwise the counterfactual will be ambiguous.

ii. But counterfactuals should not be ambiguous.

iii. We can make them unambiguous by assuming that there is a single rule, the same one all the time, about how *c* will be brought about.

iv. The rule that says "Bring c about by changing the principles that have *c* as effect to 'Set *c* = …'" is such a rule.

v. Therefore we need this rule.

vi. But this rule will not be universally applicable unless this kind of change is always possible.

vii. Therefore this kind of change must always be possible.

I have written the argument out in detail to get you to have a look at it. It is obviously fallacious. It infers from the fact that the rule in question does a needed job that it must be the rule that obtains, which is just to mistake a sufficient condition for a necessary one. So I don't think Pearl's argument will support the conclusion that changes in one principle holding fixed "everything else" are always possible and indeed are the only possibilities that matter in the evaluation of counterfactuals.

Another similar assumption that is sometimes made is that for purposes of assessing counterfactuals, changes in x_j are always presumed to be brought about by changes in z_j. But this doesn't fit with either interpretation I have given for the z's in a causal model. There is no reason that the unknown causes should be just the ones that can change; and when the z's simply serve to introduce probabilities, there isn't even a quantity there to change. To make sense of the assumption we might instead think of the z's as "exogenous" in the sense of determined outside the equations that constitute the causal model. This though will still not guarantee that they can be changed, let alone changed one at a time. Some quantities not determined by the equations of the model will nevertheless be determined by principles outside it, some may not; and some of these outside-the-model principles may be changeable and some may not.

When we consider counterfactuals for the purposes of policy and evaluation, we assume that change is really possible, change without threatening the identity of the system under study. And sometimes it is. What changes are possible and in what combinations, then, is additional information we need to put into the causal model or the causal model will not be able to tell us which counterfactuals make sense in the first place, before we begin to assess their truth and falsity.

In the economics literature Kevin Hoover makes this point explicitly.[5] Hoover distinguishes what he calls *parameters* from *variables*. Both vary, but only parameters can be changed directly -- any change the value of a variable might undergo will be the result of a change in a parameter. In formulating a causal model, then, we are to distinguish between the parameters and the variables. Moreover, each different parameter is supposed to represent a quantity that can be changed independently of every other. This implies that the quantities represented by parameters can take any combination of values in their allowed ranges; they are, formally speaking, 'variation free': $Range\ (\alpha_1,\alpha_2,\ldots,\alpha_n) = Range\ (\alpha_1)\ x\ Range(\alpha_2)\ x\ \ldots\ x\ Range\ (\alpha_n)$. We should note, though Hoover himself does not make much of this, that this is not generally the distinction intended between parameters and variables. So we must use care in taking over causal models already formulated that may distinguish parameters and variables in some other way.

1c. What is envisaged to change?

Once we have recorded what things can change, we know what counterfactuals make sense. But to assess the truth-value of any particular counterfactual we will need to know what changes are

supposed to happen. Often the exact details matter. For instance, many people feel they would not be opposed to legalizing euthanasia, *if only* it could be done in a way that would ensure that abuses would not occur.

Sometimes when we consider a policy we have a very definite idea in mind how it will be implemented. I shall call the related counterfactuals, "implementation specific". At the other end of the scale, we might have no idea at all; the counterfactuals are "implementation neutral". When we evaluate counterfactuals, we had better be clear what exactly we are presuming.

For counterfactuals that are totally implementation specific, we know exactly what we are asking when we ask "What would happen if…?"[6] For others there are a variety of different strategies we might adopt. For one, we can employ the usual devices for dealing with epistemic uncertainty. We might, for instance, assess the probabilities of the various possible methods of implementation and weight the probability of the counterfactual consequent accordingly. In the methodology of economics literature we find another alternative: Stephen LeRoy and Daniel Hausman focus on counterfactuals that would be true *regardless* of how they are implemented. I begin with LeRoy.

LeRoy's stated concern is with causal ordering among quantities, not with counterfactuals. But, it seems, he equates "*p* causes *q*" with "if *p* were to change, *q* would change as well" -- so long as we give the 'right' reading to the counterfactual. It is his proposed reading for the counterfactual that matters here. It may help to present his brief discussion of a stock

philosophical example before looking to more formal cases -- the case of birth control pills and thrombosis.

Birth control pills cause thrombosis; they also prevent pregnancy, which is itself a cause of thrombosis. LeRoy assumes that whether a woman becomes pregnant depends on both her sexual activity and whether she takes pills. Now consider: "What would happen vis-à-vis thrombosis were a particular woman to become pregnant?" That, LeRoy, points out, is ambiguous -- it depends on whether the change in pregnancy comes about because of a change in pill-taking or because of a change in sexual activity.

In his formal characterization LeRoy treats systems of linear deterministic equations. We may take these to be very sparse causal models. They are what in economics are called 'reduced form equations': "In current usage an economic model is a map from a space of exogenous variables -- agents' characteristics and resource endowments, for example -- to a space of endogenous variables -- prices and allocations."[7] The equations are expected to be functionally correct, but not to represent the causal relations among the variables, with one exception. Variables designated as 'exogenous' are supposed not to be caused by any of the remaining (endogenous) variables. Since they are functionally related to the endogenous variables, we may assume that either they are causes of some of the endogenous variables or are correlated with such causes. For LeRoy's purposes I think we must suppose they are causes.

In the context of our discussion here, with Hoover in mind, we should note one further assumption that LeRoy makes. The possible sources of change in an endogenous variable are

exactly the members of the minimal set of exogenous variables that, according to the economic model used to evaluate the counterfactuals, will fix the value of the endogenous variable. LeRoy considers a familiar supply and demand model:

(I) $q_s = \alpha_s + \alpha_{sp}p + \alpha_{sw}w$

$q_d = \alpha_d + \alpha_{dp}p + \alpha_{di}i$

$q_s = q_d = q$

Here p is price; q, quantity; w, weather; i, income. LeRoy asks what the effect of a change Δ in price would be on the equilibrium quantity. By the conventions just described, a change in price can come about through changes in weather, income or both, and nothing else. But, LeRoy, notes, "any of an infinite number of pairs of shifts in the exogenous variables 'weather' and 'income' could have caused the assumed changes in price, and these map onto different values of q."[8] Thus the question has no definite answer -- it all depends on how the change in p is brought about.

LeRoy contrasts this model with a different one:

(II) $q_s = \alpha_s + \alpha_{sw}w + \alpha_{sf}f$

$q_p = \alpha_p + \alpha_{dp}p + \alpha_{di}i$

$q_s = q_d = q,$

where f is fertilizer. Here fertilizer and weather can change the equilibrium quantity, and no matter how they do so, the change in price will be the same. In this case Leroy is content that the counterfactual, "If q were to change from Q to $Q + \Delta$,[9] p would change from $P = (Q - \alpha_p - \alpha_{di}I)/\alpha_{dp}$ to $P = (Q + \Delta - \alpha_p - \alpha_{di}I)/\alpha_{dp}$" is unambiguous (and true). The lesson he draws is the following (where I substitute counterfactual language for his causal language): "[Counterfactual]

statements involving endogenous variables as [antecedents] are ambiguous except when all the interventions consistent with a given change in the [antecedent] map onto the same change in the [consequent]."[10] I think the statement as it stands is too strong. Some counterfactuals are, after all, either implicitly or explicitly implementation specific. What LeRoy offers is a semantics for counterfactuals that are, either implicitly or explicitly, implementation neutral. In this case the consequent should obtain *no matter what possible change occurs to bring the antecedent about*.

Dan Hausman seems to have distinguished between implementation-specific and implementation-neutral counterfactuals, too, as I do here, though I do not think he explicitly says so. He considers an example in which engineers designing a nuclear power plant ask, "What would happen if the steam pipe were to burst?"[11] The answer, he argues, depends on how it will burst. "Responsible engineers", he argues, must look to the origins of the burst "when the consequences of the pipe's bursting depend on what caused it to burst."[12]

On the other hand, when Hausman turns to providing some constraints that a possible-world semantics for counterfactuals must satisfy, he seems to be concerned with implementation-neutral counterfactuals. The results are similar to LeRoy's: any semantics that satisfies Hausman's constraints should give the same result as LeRoy's prescription when restricted to counterfactuals evaluated via what LeRoy calls an 'economic model'. The Hausman constraint on the similarity relation between possible worlds that matters to our discussion here is

> ***SIM 2***. *It doesn't matter which cause is responsible.* For any event *b*, if *a* and *c* are any two causes of *b* that are causally and counterfactually independent of one another,

there will be non-*b* possible worlds in which *a* does not occur and *c* does occur that are just as close to the actual world as are any non-*b* possible worlds with *a* and without *c*, and there will be non-*b* possible worlds without *a* and with *c* that are just as close to the actual world as are any non-*b* possible worlds without both *a* and *c*.[13]

Look back at LeRoy's model (I) for illustration, where weather and income are the causes by which either price or quantity can change. It is easiest to see the results if we first solve for *p* and *q*:

$$q = (\alpha_{dp}\alpha_s - \alpha_{sp}\alpha_d + \alpha_{dp}\alpha_{sw}w - \alpha_{sp}\alpha_{di}i)/(\alpha_{dp} - \alpha_{sp})$$

$$p = (\alpha_s - \alpha_d + \alpha_{sw}w - \alpha_{di}i)/(\alpha_{dp} - \alpha_{sp})$$

If p changes by ΔP with *w* fixed, then *i* must have changed by $\Delta P(\alpha_{sp} - \alpha_{dp})/\alpha_{di}$ and so q will change by $\Delta Q = \alpha_{sp}\Delta P_i$. If on the other hand *i* is fixed, then w must have changed by $\Delta W = \Delta P(\alpha_{dp} - \alpha_{sp})/\alpha_{sw}$ and so $\Delta Q = \alpha_{dp}\Delta P$. Now we can bring in SIM 2). If *q* changes ('*q*' is here the analogue of '*b*' in SIM 2) some world in which w (the analogue of '*a*') changes will be just as close as any world in which *i* (the analogue of '*c*') changes. But the world in which *w* changes and *i* stays fixed and the world in which *i* changes and *w* stays fixed have different values for the change in *q*. Yet they are equally close. So the truth value of counterfactual claims about what would happen to *q* were *p* to change by ΔP are undefined.

So we may have counterfactuals that are implementation specific; we may have ones that assume some one or another of a range of possible implementations; and we may have implementation-neutral ones where we wish to find out what would happen no matter how the change in the

antecedent is brought about. For thinking about policy we had better know which kind of counterfactual we are asserting and ensure that our semantics is appropriate to it.

2. Impostor counterfactuals

The kinds of "What if…?" questions we ask in planning and evaluating are in sharp contrast with a different kind of 'counterfactual' that occupies economists as well -- the impostor counterfactuals. Like the counterfactuals I have so far been discussing these too are evaluated relative to a causal model. But they are not used directly in planning and evaluation. Rather they are used to define certain causal concepts. For Heckman the relevant concept is *causal effect*; for LeRoy, *causal order*. I shall discuss LeRoy first.

2a. Le Roy

I have urged that in order to assess counterfactuals, we need a causal model. Recall that LeRoy begins with a sparse causal model: a reduced form equation that links the endogenous variables to a set of exogenous variables, where he supposes that no exogenous variables are caused by any endogenous ones and that the exogenous variables completely determine the values of the endogenous variables.[14] The task is to say something about the causal order of the endogenous variables and, I take it, about the strength of influence of one on another. Let Z_j be the minimal set of exogenous variables that determine x_j and define Z_{ji} as Z_j - Z_i. Then x_c causes x_e if and only if there is a (scalar) γ_{ec} and a (vector) δ_{ec} such that

$$x_e = \gamma_{ec} x_c + \delta_{ec} \overline{Z}_{ec}.$$

This means that x_e is determined completely by x_c plus a set of exogenous variables that do not participate in determining x_e; that is, there is no z that both helps fix the first term in the above equation and also helps fix the second.

What what-if question does γ_{ec} answer? It answers an implementation-neutral counterfactual: by how much would x_e change were x_c to change by a given amount, no matter how the change in x_c is brought about? This is often an important question for us to be able to answer; it may also important to know for the system we are dealing with that it has no answer: there is nothing general, or implementation neutral, that we can say; how much the effect changes cannot be calculated without knowing what the method of implementation will be.

There are two points I would like to make about LeRoy's approach. First I admit that these counterfactuals are in no way 'impostors' -- they ask genuine what-if question whose answers we frequently need to know. Nevertheless they are severely restricted in their range of application. For vast numbers of systems the answer to LeRoy's counterfactual question will be that it has no answer: there is no implementation-neutral change that would occur in the effect consequent on a change in the cause.

Second, LeRoy's definition answers one very special kind of causal question -- it asks about how much, if one factor changes in any way whatsoever, a second factor will change. But it does not answer the question of how much one factor *contributes* to another. For a simple example where

the two questions have different answers consider a system governed by the following two causal laws:

(CM1) $q\ c = \alpha_{qz} z$

$p\ c = \alpha_{pz} z.$

Compare this with a system governed by different laws

(CM2) $p = \alpha_{pz} z$

$q = \alpha_{qp} p.$

It should at least in principle be possible for two such systems to exist. The two systems have different causal structures and different answers to the question, "How much does p contribute causally to q?" In the second system the answer is given by α_{qp}. In the first the answer is "nothing". Yet in cases where $\alpha_{qz} = \alpha_{qp}\alpha_{pz}$ there will be exactly the same answer to LeRoy's counterfactual question: If p were to change by Δp, no matter how it does so q would change by $\alpha_{qz}\Delta p = \alpha_{qp}\alpha_{pz}\Delta p$.

In my view we have a large variety of causal concepts, applicable to a variety of different kinds of systems in different situations, and there is also a large variety of different kinds of causal and counterfactual questions we can ask, many of which only make sense in particular kinds of systems in particular circumstances. LeRoy asks a specific, explicitly articulated counterfactual question, and I take it that that is all to the good. We must be careful, however, not to be misled by his own use of the language of "causal order" to suppose it tells us whether and how much one quantity causally contributes to another.

2b. Heckman

Heckman also uses counterfactuals to answer what he labels as causal questions. I find his usage of them less transparent. Like LeRoy, Heckman asks an explicit, well articulated counterfactual question, in his case an implementation-specific question. Again, as with LeRoy, the question has an answer only in certain very restricted systems -- essentially, as I shall explain, in Galilean-style experiments. As far as I can see, the primary interest in Heckman's counterfactuals is that they serve as a tool for answering a non-counterfactual question, a question about causal contributions. But questions about causal contributions can be asked -- and answered -- for situations that are not Galilean experiments, where the counterfactuals Heckman introduces do not make sense. This is why I say that they are impostors. They seem to be the issue of interest; they are certainly the *topic*. But in fact they are only a tool for answering a different question -- a causal question -- and at that, for answering that question only in very restricted kinds of systems, kinds that are not generally the ones of concern.

Before we turn to Heckman it may be helpful to begin with work that will be more familiar to philosophers, from the book *Causality* by Judea Pearl. Pearl gives a precise and detailed semantics for counterfactuals. But what is the semantics a semantics of? What kinds of counterfactuals will it treat, used in what kinds of contexts? Since Pearl introduces them without comment we might think that he has in mind natural language counterfactuals. But he presents only a single semantics with no context dependence, which does not fit with natural language usage.

Worse, the particular semantics Pearl develops is unsuited to a host of natural language uses of counterfactuals, especially those for planning and evaluation of the kind I have been discussing. That is because of the very special way in which he imagines that the counterfactual antecedent will be brought about: by a precise incision that changes exactly the counterfactual antecedent and nothing else (except what follow causally from just that difference). But when we consider implementing a policy, this is not at all the question we need to ask. For policy and evaluation we want generally to know what would happen were the policy really set in place. And whatever we know about how it might be put in place, the one thing we can usually be sure of is that it will not be by a precise incision of the kind Pearl assumes.

Consider for example Pearl's axiom of composition, which Pearl proves to hold in all causal models, given his characterization of a causal model and his semantics for counterfactuals. This axiom states that "if we force a variable (W) to a value *w* that it would have had, without our intervention, then the intervention will have no effect on other variables in the system."[15] This axiom is reasonable if we envisage implementations that bring about the antecedent of the counterfactual in as minimal a way as possible. But it is clearly violated in a great many realistic cases. Often we have no idea whether the antecedent will in fact obtain or not, and this is true even if we allow that the governing principles are deterministic. We implement a policy to ensure it will obtain -- and the policy may affect a host of changes in other variables in the system, some envisaged and some not.

We should note that the same problem arises for Lewis-style semantics. If the antecedent of a counterfactual obtains, then our world, with things as they actually happen in it, is the nearest

possible world for evaluating the truth value of the counterfactual. There is no room then for anything to change as a result of the antecedent being implemented.[16]

Heckman, unlike Pearl and Lewis, is keen that causal models *model* how change is brought about. So in defining causal efficacy he does not adopt Pearl's semantics in which laws are changed *deus ex machina*. But he does adopt a similar device. Pearl limits his causal definitions to systems in which the principles responsible for a given factor, with all their causes, can be changed to produce any required value for that factor, without changing any other principles or other "initial" values. Heckman limits his definitions to causal principles in which the causes are variation free. This means that if only the system runs 'long enough', the effect (intended as the antecedent of the counterfactual) will naturally take any required value, while the remaining causes, all other principles, and all other initial values stay the same. The counterfactual change in an antecedent with 'everything else' the same will 'eventually' be factual. Heckman stresses, thus, that what matters for his definitions is natural variability within the system, not changes in the principles under which it operates.

Heckman begins his treatment with *causal functions*. These govern very special kinds of causal systems, systems that mimic experiments: "Causal functions are … derived from conceptual experiments where exogenously specified generating variables are varied….The specification of these hypothetical variations is a crucial part of model specification and lies at the heart of any rigorous definition of causality."[18]

Heckman tells us three things about causal functions: i) They "describe how each possible vector of generating variables is mapped into a resulting outcome", where the generating variables "completely determine" the outcome.[19] ii) They "derive from" -- or better, I think, 'describe' -- conceptual experiments. iii) Touching on questions of realism and of model choice, models involving causal functions are always underdetermined by evidence; hence, as Heckman sees it, causality is just "in the head" since the models relative to which it is defined are just in the head. From this I take it that causal functions represent (a probably proper subset of) the causal principles under which these special experiment-like systems operate, where the right-hand-side variables -- the ones Heckman calls the "generating variables" -- form a minimal complete set of causes of the quantity represented on the left[19] and where each cause can vary independently of the others.

Imagine that the causal function for an outcome y is given by

$y = g(x_1, \ldots, x_n)$.

We can now define the *causal* or *counterfactual effect* of x_j on y fixing the remaining factors in the causal function (Heckman seems to use the terms 'causal effect' and 'counterfactual effect' interchangeably):

(**Causal effect of x_j on y**)

$[\Delta y / \Delta x_j = x_j' - x_j''] =_{df} g(x_1, \ldots, x_j', \ldots x_n) - g(x_1, \ldots, x_j'', \ldots x_n)$.

As Heckman insists, in order for this definition "to be meaningful requires that the x_j can be independently varied when the other variables are fixed so that there are no functional restrictions connecting the arguments….it is thus required that these variables be variation-

free".[20] I shall call the counterfactual effect as thus defined a *Galilean counterfactual* since, as I remarked, it is just the kind effect we look for in a Galilean experiment.

I should note that Heckman himself treats of double counterfactuals since the outcome variables he discusses are often themselves counterfactuals: y_0 is the value a given quantity would take were a specified 'treatment' to occur; y_1, the value it would take were the treatment not to occur. These values, he supposes, are fixed by deterministic causal functions. Relative to these causal functions we can then ask about the causal efficacy of a certain quantity -- including the treatment itself -- on the counterfactual quantities y_0 and y_1. So we can consider, for example, what difference a change in social security regulations would have on the amount of savings that would obtain if there were a tax cut versus the difference the change would make were there no tax cut. I will not be concerned with these double-barreled counterfactuals here. They do not appear in Heckman's discussion of the supply and demand equations, which will suffice as illustrations of my central point.

Heckman considers simultaneous supply and demand equations. For simplicity we can look at the specific equations that we have already considered above, where I have added the additional equilibrium constraint on price:

(I') $q_s = \alpha_s + \alpha_{sp}p_s + \alpha_{sw}w$

$q_d = \alpha_d + \alpha_{dp}p_d + \alpha_{di}i$

$q_s = q_d = q$

$p_s = p_d = p.$

Heckman points out that these equations do not fit Pearl's scheme since they are not recursive and hence Pearl's method for assessing counterfactuals will not apply. This fits with familiar remarks about these kinds of systems: p and q are determined jointly by exogenous factors. It seems then that it makes no sense to ask about how much a change in p will affect a change in q. To the contrary, Heckman points out: we can still assess causal efficacy using his definition -- so long as certain 'exclusion' conditions are met.

Say we want to assess the causal/counterfactual effect of demand price on quantity demanded. We first look to the reduced form equations

$$q = (z_d, z_s)$$

$$p = (z_d, z_s)$$

where z_d is the vector of exogenous variables in the demand equations and z_s, those in the supply equations. In LeRoy's equations (I'), $z_d = i$ and $z_s = w$. Heckman takes these to be causal functions, otherwise the causal model has not properly specified the 'exogenous' variables. That means that the exogenous variables are 'generating variables' for p and q and that they are variation free. Now the task is easy: "Assuming that some components of [z_d] do not appear in [z_s], that some components of [z_s] do not appear in [z_d], and that those components have a non-zero impact on price, one can use the variation in the excluded variables to vary [p_d or p_s in the reduced form equations] while holding the other arguments of those equations fixed."[21] The result (using the equality of p_d and p_s and of q_d and q_s) is

$$\partial q_d / \partial p_d = (\partial q / \partial z_s(e)) / (\partial p / \partial z_s(e))$$

where $z_s(e)$ is a variable in z_s that is excluded from z_d and that, as he puts it, "has an impact on" p_d. In (I') this job can be done by w; the causal effect thus calculated of p_d on q_d is α_{dp}.

Notice how much causality is involved here. By definition we are supposed to be evaluating the change in q_d holding fixed all the factors in a causal function for q_d except p_d. What we actually do is hold fixed z_d while z_s varies. Presumably this is okay because z_s is a cause of p_d that can produce variations in p_d while z_d is fixed; and z_d being fixed matters because z_d constitutes, along with p_d, a minimal full set of causes of q_d. So when the exclusion condition is satisfied, the demand equation is a causal function and the counterfactual definition of causal effect is meaningful.

Now consider a slightly altered set of equations:

(I'') $q_s = \alpha_s + \alpha_{sp}p_s + \alpha_{sw}w + \alpha_{si}i$

$q_d = \alpha_d + \alpha_{dp}p_d + \alpha_{di}i + \alpha_{dw}w$

$q_s = q_d = q$

$p_s = p_d = p.$

Now the demand equation cannot be treated as a causal function and the question of the causal effect of demand price on quantity demanded is meaningless. This is true despite the fact that α_{dp} still appears in the equation and it still represents something -- something much the same one would suppose -- about the bearing of p_d on q_d. The intermediate case seems even stranger. Imagine that $\alpha_{sw} = 0$. Now α_{sp} measures a counterfactual effect but α_{dp} does not.

2c. Cartwright

I have an alternative. But I should note that I have a stake in this discussion since I have been stressing the importance of independent variability for over 15 years; I just think it plays a different role than Heckman (and Pearl and Hausman and Woodward) ascribe to it.

I begin with causal principles. At this level of discussion I myself am a realist about the principles of our causal models: they are correct if and only if they approximate well enough to the causal laws that govern the operation of the system in question. Heckman, it seems, is not a realist. But that does not matter here since he himself has introduced the notion of a causal function. A causal principle is just like a causal function but without the restriction that the causes (or "generating variables") are variation free. I shall continue to restrict attention to linear causal models. Then, for a given causal model, *the contribution a cause* x_c *makes to an effect* x_e is just the coefficient of x_c in any causal principle for x_e in the model.[22] It is easy to show for linear models that where Heckman's measure for the causal/counterfactual effect of x_c on x_e applies, it will have the same value as the contribution x_c makes to x_e.

Given this characterization we see that the contribution of p_d to q_d is the same in (I') and (I''). What is different is that in (I') we have a particular way to find out about it that is not available in (I''). (I') is what I have called *an epistemically convenient system*:[23] it is a system in which we can find out what a cause, x_c, contributes to an effect, x_e, in one particular simple way: hold fixed all the other contributions that add up to make the effect the size it is; then vary the cause and see how much x_e varies. Any difference has to be exactly the contribution that x_c adds. This does not mean, however, that for systems where this independent variation is not possible, all is lost.

There are hosts of other legitimate ways of defending claims about the size of causal contributions that apply both in systems with independent variation and in ones without.[24]

3. Epistemic convenience versus external validity

I began my discussion with reference to impostor counterfactuals. There is a sense in which the counterfactual questions that Heckman focuses on are genuine: if we are talking about the right kinds of systems -- epistemically convenient ones -- they ask genuine implementation-specific what-if questions. But there are two problems. First, few systems we confront are epistemically convenient. The vast majority are not. For these Heckman's measures are irrelevant.

Second, even if we are studying an epistemically convenient system there is a puzzle about why we should wish to ask just these implementation-specific questions. If we were thinking of setting policy or evaluating the success of some program in the system, then these, with their very special method of implementation, might be relevant sometimes. But there is no necessity to implement policies in the single way highlighted by Heckman; generally we would want to consider a variety of different methods of implementation and frequently to assess implementation-neutral counterfactuals as well. Even in epistemically convenient systems, the Galilean counterfactuals that Heckman studies often have no privileged role.

There are two familiar enterprises where they do have a special role. The first is in trying to determine if, and to what degree, one factor contributes causally to another. In an epistemically

convenient system we can ask Galilean-type counterfactual questions; and the answers we obtain will double as answers to our causal questions. They are a tool for finding out answers to our causal questions. But note that they are only a tool for finding out about causes in our special epistemically convenient systems. For other systems we cannot even ask these counterfactual questions, let alone let the answers to them supply our causal answers as well.

The other is in Heckman's own field, evaluation. In setting up new programs, we might try to set them up in such a way that the causal contribution they make to the result can be readily disentangled from the contribution of other factors. Of particular concern are other factors that might both contribute to the effect independently of the program and also make it more likely that an individual entered (or failed to enter) the program. If we can arrange the setup of our program so that it is epistemically convenient, then again we can answer Galilean counterfactual questions -- "What difference would there be in outcome with the program present versus the program absent, holding fixed all other contributions to the outcome?" And again these counterfactual questions will tell us the contribution the program makes, since in these circumstances the difference in outcome between when the program is present and when it is absent must be exactly the contribution the program makes. So we can use information about Galilean counterfactuals to learn about the causal contributions of the program we set up. Still, all we learn is about that program in those special epistemically convenient circumstances.

In either case, whether it be experimental systems or program set-ups that we engineer to make the measurement of causal contributions easy, we need to ask, why should we be interested in causal contributions in these very special -- and rare -- kinds of systems? The answer is clear.

Generally we want this information because it will tell us something about causal contributions in other systems as well. But we confront here the familiar problem of internal and external validity. In an epistemically convenient (linear) system, using counterfactual differences as a measure of causal contributions is provably valid: internal to the situation this method is bound to give us correct results about the question of interest. But nothing said in this discussion bears on external validity: when will the results that we can be sure are correct in a convenient system hold elsewhere?

Sometimes this issue is discussed in the economics methodology literature under the heading 'invariance'. This is often with something like equation set (I') in mind. Here we can find out the causal contribution, α_{dp}, of p_d to q_d by calculating the difference in Galilean counterfactuals as p_d changes via w holding fixed i. Then we might imagine that everything already in place about the causal principle for q_d would stay the same even if weather became an influence on quantity demanded. Thus we suppose that the second equation can be replaced with

$$q_d = \alpha_d + \alpha_{dp}p_d + \alpha_{di}i + \alpha_{dw}w.$$

We then say that the equation for q_d remains *invariant* as α_{dw} changes from zero to non-zero, or possibly we suppose it invariant over any range of values for α_{dw}. This though is only one kind of assumption we might make about the use to which we can put the information we learn about the causal contribution that one factor makes to another. Since I have written at length about this topic elsewhere,[25] I will not pursue it further here.

There are two points that matter to my argument here. The first is that assumptions about where this information can be put to use are not justified by anything we have discussed so far, and in

particular not by any information about counterfactuals of the kinds I have explored. Showing that results on causal contributions have external validity -- and how far and of what kind -- requires a different methodology altogether.

Second, when we export the information gleaned from Galilean counterfactuals in epistemically convenient systems elsewhere, it is not as information about counterfactuals but rather as information about causal contributions. In most systems to which we will carry our results, Galilean counterfactual questions do not even make sense. This supports my claim that both as counterfactuals and as causal surrogates, Galilean counterfactuals are impostors. They do not carry over as counterfactuals to non-epistemically convenient systems; and in epistemically convenient ones they are usually of interest, not on their own as genuine what-if hypotheses but only as tools for measuring causal contributions. Even then the results about causal contributions are of use outside the highly restricted systems in which they are established only if specific assumptions about the external validity of the results are warranted.

4. Causal decision-theory

As another illustration of the conflation of Galilean counterfactuals with more realistic implementation-specific ones, consider causal decision-theory. Various versions of causal decision-theory made the same mistake I am pointing to, but in reverse: The aim was to evaluate genuine counterfactuals but we ended up with a measure that measured the causal contribution of

a factor and not the counterfactual effects of the factor being implemented. Let us consider a very simple case.

Given my fear of lung cancer, should I quit smoking? Presumably the answer is "yes" if the expected utility if I were to quit is greater than if I were to continue; or

Counterfactual decision formula:

$P(S\Box\rightarrow L)\ U(S\&L) + P(S\Box\rightarrow\neg L)\ U(S\&\neg L) <\ P(\neg S\Box\rightarrow L)\ U(\neg S\&L) + P(\neg S\Box\rightarrow\neg L)\ U(\neg S\&\neg L)$

where *S = I smoke*, *L= I get lung cancer*, *U(X)* = utility of *X*, and where I shall assume the probabilities are personal probabilities read off from the population probabilities.

Conventionally in decision theory *P(B/A)* appeared in this formula instead of $P(A\Box\rightarrow B)$:

'Conventional' decision formula:

$P(L/S)\ U(S\&L) + P(\neg L/S)\ U(S\&\neg L) <\ P(L/\neg S)\ U(\neg S\&L) + P(\neg L/S)\ U(\neg S\&\neg L)$

but it became apparent that this would not do. As the slogan has it: The probability of a counterfactual conditional is not a conditional probability. I can illustrate why with a caricature of a hypothesis mooted by R.A. Fisher. Perhaps smoking does not cause lung cancer; rather the observed probabilistic dependence of lung cancer on smoking arises entirely because both are the result of some gene that is prevalent in the population. Then it might well be the case that $P(L/S)>>P(S/\neg L)$, but it would not make sense to give up smoking if one loved it in order to avoid lung cancer. To keep the example simple I shall suppose that there is no other cause of lung cancer besides the two possible causes, smoking and the gene.

Since on the 'Fisher' hypothesis the probabilistic dependence between S and L is due entirely to the fact that each is itself dependent on the gene, the dependence between them should disappear if we condition on the presence or absence of the gene. This led causal decision theorists to substitute the partial conditional probability $P(L/\pm S\pm G)$ for $P(L/\pm S)$, depending on whether I do indeed have the gene or not (G = *I have the smoking/lung cancer gene*). If, as we might expect, I have no idea at all whether I have the gene, then I should average over $P(L/\pm S\pm G)$, where the weights for the average would reasonably be based on the frequency with which G appears in the population: $P(+G)$, $P(\neg G)$. In case we can make the additional assumption that the only bearing that the gene has on my utility is through smoking and lung cancer,[26] this line of reasoning results in

Causal decision formula:

$$[P(L/S\&G)\,P(G) + P(L/S\&\neg G)\,P(\neg G)]\,U(S\&L) + [P(\neg L/S\&G)\,P(G) + P(\neg L/S\&\neg G)\,P(\neg G)]\,U(S\&\neg L) < [P(L/\neg S\&G)\,P(G) + P(L/\neg S\&\neg G)\,P(\neg G)]\,U(\neg S\&L) + [P(\neg L/\neg S\&G)\,P(G) + P(\neg L/\neg S\&\neg G)\,P(\neg G)]\,U(\neg S\&\neg L)$$ [27]

In the case when G is independent of S ($P(\pm G/\pm S) = P(\pm G)$), this formula reduces to the 'conventional' formula.

Notice that the difference $P([S\Box\!\rightarrow L]/\pm G) - P([\neg S\Box\!\rightarrow L]/\pm G)$ is given by $P(L/S\&\pm G)\,P(\pm G) - P(L/\neg S\&\pm G)\,P(\pm G)$. This latter formula is a direct analogue to Heckman's formula for the causal/counterfactual difference for values: Hold fixed the other causes of the effect in question and see what difference occurs when the targeted cause varies on its own; only in this case we look not to the difference in values of the effect as the cause varies but rather to the difference in probabilities. I shall by extension call this the *probabilistic causal/counterfactual difference*. It is

clearly not defined if *S* and *G* are not variation-free; when it is defined and they are variation free, we can also by analogy take the formula to provide a measure of the *probabilistic causal contribution* of *S* to *L* given *G* or given ¬*G*.[28]

Like the value-based causal/counterfactual difference this too is more like the counterfactual difference we look for in a Galilean experiment than the implementation-specific difference that might occur in real cases. The particular example chosen tends to obscure this point (as did many others focused on in the early days of causal decision theory). In our case we have only one other cause on the tapis and it is unlikely to be changed by any method by which we might come to stop smoking. But suppose that the way in which I will be brought, or bring myself, to stop smoking has some chance of altering whether I have the relevant gene or not. In that case, if we assume that the causal contributions of separate factors are additive, a better formula for the implementation-specific probabilistic counterfactual difference might be[29] (letting *cc(A,B/C)* stand for the causal contribution of *A* to *B* in the presence of *C*):

$$P([S\Box\!\rightarrow\!L]/\pm G) - P([\neg S\Box\!\rightarrow\!L]/\pm G) = cc(S,L/\neg G)\, P([S\Box\!\rightarrow\!\neg G]/\pm G) +$$

$$[cc(S,L/G) + cc(G,L/S)\, P([S\Box\!\rightarrow\!G]/\pm G).$$

I offer this formula as an illustration to make a specific point. Behind the story is a small causal model based on the little story I told about smoking, the gene and lung cancer plus the assumption that contributions from separate causes combine additively. And that buys us some advance. But it does not eliminate the counterfactuals altogether. We still need a model involving the implementation variables and the relation to the system to calculate the probability of the remaining counterfactuals. The second model in cases like this will often be far more ad hoc and involve far more local knowledge than the one that models the basic system itself.

The overall point of this discussion, however, is that causal decision theories typically employ a measure that depends entirely on the causal contribution of the action in question. But what is needed, as in policy deliberations in general, is a formula that involves implementation-specific counterfactuals across the range of implementations that might in fact obtain -- i.e., 'genuine' counterfactuals.

5. Conclusion

I have claimed that the impostor counterfactuals of current interest in economics provide a tool to measure causal contributions, though a tool limited in its use to Galilean experiments. It is important to stress that questions about causal contributions are central questions that definitely need answering for the kinds of systems we live in and use. I began with genuine counterfactuals. For purposes of planning and evaluation we need answers to genuine what-if questions, both implementation-specific questions and implementation neutral ones. But I have now come full circle. We cannot evaluate the counterfactuals unless we have a causal model. And what is a causal model, in the context of answering genuine what-if questions? A causal model is a set of causal principles that represent our hypotheses about just the causal issue I describe: to what degree does one factor contribute causally to another. This is the information we need for genuine counterfactuals, and impostors play at best a very indirect role in helping to provide it.

References

Cartwright, N. 1989. *Natures Capacities and their Measurement*. Oxford: Clarendon Press.

------ 1999. *The Dappled World: A Study of the Boundaries of Science*. Cambridge: Cambridge University Press.

------ 2002. "Against Modularity, the Causal Markov Condition and Any Link Between the Two: Comments on Hausman and Woodward." *British Journal for the Philosophy of Science* 53: 411-453.

------ 2003. "Two Theorems on Invariance and Causality." *Philosophy of Science* 70: 203-224.

------ forthcoming. "From Metaphysics to Method: Comments on Manipulability and the Causal Markov Condition."

Hausman, D. 1998. *Causal Asymmetries*, Cambridge University Press, Cambridge.

Hausman, D., and J. Woodward. 1999. "Independence, Invariance, and the Causal Markov Condition." *British Journal for the Philosophy of Science* 50: 521-583.

------ 2004. "Modularity and the Causal Markov Condition: A Restatement." *British Journal for the Philosophy of Science* 55: 147-161.

Heckman, J. 2001. "Econometrics, Counterfactuals and Causal Models." Keynote Address *International Statistical Institute*. Seoul, Korea.

Hoover, K. 2001. "Causality in Macroeconomics", Cambridge: Cambridge University Press.

LeRoy, S. 2003. “Causality in Economics” MS, University of California, Santa Barbara.

Pearl, J. 2000 “Causality: Models, Reasoning and Inference”. Cambridge: Cambridge University Press.

Reiss, J., and N. Cartwright. 2003. “Uncertainty in Econometrics: Evaluating Policy Counterfactuals” *Causality: Metaphysics and Methods Technical Report*. Centre for the Philosophy of Natural and Social Science, LSE CTR 11-03.

Notes

Research for this paper was supported by an AHRB grant, *Causality: Metaphysics and Methods*, and by a grant from the Latsis Foundation. I am grateful to both. Many of the ideas were developed jointly with Julian Reiss (see our paper “Uncertainty in Econometrics: Evaluating Policy Counterfactuals”) ; I also want to thank him.

1. The system is functionally transitive iff $x_k\, c = f(...x_j...)$ and $x_j\, c = g(...) \rightarrow x_k\, c = f(...g(...)...)$.
2. Hausman, 1998, 119.
3. Or, more carefully, our confidence in a counterfactual can be no higher than our confidence that our casual model will produce correct predictions about this counterfactual.
4. In Causal Asymmetries Hausman aims to make this connection. But, as the title suggests, generally what he succeeds in doing is using his claims to obtain causal order. For instance, he shows that, given his claims about the independent variability of causal

principles, if b counterfactually depends on a, then a causes b. This is an important result. But to establish it requires the prior assumption that if a and b are counterfactually connected then either a causes b or the reverse or the two have a common cause plus his own (as opposed for instance to David Lewis's) constraints on the nearness relation for a possible-world semantics for counterfactuals (which I describe below in discussing implementation-neutral counterfactuals). Hausman and Woodward 1999 also claim that the independent variability assumption implies the causal Markov condition. But they do not show that the assumption implies the causal Markov condition, which is false, but rather that there are some systems of equations in which both are true and that it is, roughly speaking, "the same" features of these systems that guarantee both assumptions (see Cartwright 2002 and forthcoming.)

5. Hoover, 2001.

6. Or rather, we know this relative to the factors included in the causal model. Presumably no causal model will be complete, so this remains as a source of ambiguity in our counterfactual claims.

7. LeRoy, 2003, 1.

8. LeRoy, 2003, 6.

9. I shall follow LeRoy's convention throughout and use lower-case letters for variables and upper case for their values.

10. LeRoy, 2003, 6.

11. Hausman, 1998, 122.

12. *Ibid.*

13. Hausman 1998, p 133.

14. Note that the reduced form equation need not be a causal function in the sense that I shall introduce from Heckman, since LeRoy allows that the external variables may not be variation free, though he thinks it would be odd if they were not.

15. Pearl, 2000, 229.

16. For a longer discussion of Pearl and Lewis see Reiss and Cartwright 2003.

17. Heckman, 2001, 14.

18. Heckman, 2001, 12.

19. Or, keeping in mind Heckman's view that causality is only relative to a model, the right-hand-side variables record what the model designates as causes.

20. Heckman, 2001, 18.

21. Heckman, 2001, 36.

22. Recall that the discussion here is limited to linear systems; the concept of a causal contribution is more complex in non-linear systems. Also note that this supposes that all principles in the model with x_c on the right-hand-side and x_e on the left will have the same coefficient. This will be the case given a proper statement of 'transitivity' and the definitions for the form of causal principles sketched in Cartwright 2003.

23. For a definition see Cartwright 2003.

24. For further discussion see my 1989. It should be admitted of course that once the causes need not be variation free, the simple operational way of defining causal contribution in a way analogous to Heckman's definition of causal/counterfactual effect is not available. But, as we know, there are compelling arguments in the philosophical literature to establish that demanding operational definitions is both too strong and too weak a

requirement -- it lets in concepts that do not make sense and does not provide a proper understanding of those that do.

25. Cartwright 1989, 1999.

26. So that $U(\pm S \pm L \pm G) = U(\pm S \pm L)$.

27. When there is more than one common cause involved, the usual generalization of this formula conditions on the state descriptions over the common causes, weighted with the probabilities with which each state description obtains.

28. In the linear models assumed in section 2., the coefficients of each variable are assumed to be functionally independent of the values of all variables so relativization analogous to the relativization to $+G$ and $\neg G$ here was not necessary. The assumption here analogous to that in section 2. would be that S's contribution to L is the same in the presence and in the absence of G.

29. I offer this as a plausible example. Whether it is the 'correct' formula or not will, as I have argued, depend on the details of the causal model; and, as I have also already noted, we do not yet have very good prescriptions for getting from the great variety of different kinds of models we employ to methods of evaluating the various different kinds of implementation-neutral and implementation-specific counterfactuals we may need for policy.