Actual Causation: A Stone Soup Essay

Draft of February 9, 2007

Authors so far, if they are willing

David Danks[1,2], Frederick Eberhardt[1], Bruce Glymour[3], Clark Glymour,[1,2] Joseph Ramsey[1], Richard Scheines[1], Peter Spirtes[1], Choh-Man Teng[2] and Jiji Zhang[4i]

Abstract

We argue that current discussions of criteria for actual causation are ill-posed in several respects. (1) The methodology of current discussions is by induction from intuitions about an infinitesimal fraction of the possible examples and counterexamples; (2) novel problem cases occur when more variables are considered; (3) as deployed in discussions of actual causation, "neuron diagrams" and causal Bayes net diagrams are ambiguous; (4) actual causation typically involves changes in a system state that produce changes, and thus, unlike most current proposals, is commonly relative to an initial system state; (5) a well-developed principled theory accounting for actual causes that are changes that produce changes has been available for many years using causal Bayes nets; (6) there is no reason to think that philosophical judgements about cases are normative, and (7) there is a dearth of relevant psychological research that bears on whether various philosophical accounts are descriptive.

Once upon a time a wanderer came into a hungry village. He filled an iron cauldron with water, and built a fire under it. Then he dropped an ordinary looking stone into the water. "I do like a tasty stone soup" he announced. Soon a villager added a cabbage to the pot. Another added some salt beef . And soon other villagers added potatoes, onions, carrots, mushrooms, and so on, until there was a meal for all.

The question of when one event or circumstance causes another has been the subject of two recent collections of philosophical essays, (Dowe and Noordhof, 2004; Collins, et al, 2004), of a lengthy chapter in a prize-winning book (Woodward, 2003), of a connected pair of articles amounting to a short book (Halpern and Pearl, 2005a, 2005b), as well as of several other recent articles (Gilles, 2005; Spohn, 2005). Most of the literature is roughly Socratic and inductive: analyses are considered and a handful of "intuitive" story examples are produced as evidence, followed by counterexamples and other proposed analyses. Some formal structure has been imposed by reconstructing stories as Bayes net causal models: directed acyclic graphs (DAGs) whose vertices are variables and whose directed edges mark functional dependencies—truth functions or other deterministic relations, or conditional probability relations. A "causal model" then consists of a graph and a set of appropriate functional dependencies; a state is an assignment of values to the variables, and counterfactual claims refer to the results of exogenous interventions in the system.[ii]

Using the counts these representations permit, we argue that (i) the Socratic strategy for finding or testing a characterization of actual causation by intuitions about causal Bayes net cases is futile because the number of cases potentially presenting distinct challenges to theories is unsurveyably large even with small numbers of potential causes. We also argue that (ii) a restriction on the relation between graphs and truth functions, which we call the "test pair' condition, is presumed in all accounts of actual causation for causal models, but (iii) does not significantly reduce the number of cases. We consider symmetry principles that result in partitions that reduce the number of distinct causal models, but we note (iv) they are insufficient to compensate for the super-exponential growth in cases as the number of potential causes increases. Using an example with five binary causes and one effect, we argue (v) that novel structures that distinguish among proposals for actual causation arise with five potential causes, so that it is not plausible that all problems of interest are realized by cases with three or fewer potential causes. We further argue (vi) that the common graphical representation of actual causation is systematically ambiguous, and (vii) unambiguous actual causal relations concerning changes from one initial state to another consequent on interventions can be represented unambiguously by causal Bayes nets, and require no induction over a vast sea of cases. Finally, we note, with regret, (viii) that present philosophical theories of actual causation are justified only by intuitions about particular cases of authors and their circles of acquaintances, with the result that even in very simple cases philosophical theories may differ from the consensus judgements of informed lay people—a matter about which there is almost no data.

2. Counting Graphs and Truth Functions

Many of the deterministic examples in discussions of actual causation implicitly presuppose formal structures of the following kind:

1. Events are represented by variables (usually taking two values but in principle without limit), with one value (e.g., "0") possibly marked for absences.
2. A directed acyclic graph with the variables as vertices.
3. *Laws* are given by deterministic or stochastic functions specifying values for each variable as a function of the values of its parents in the graph. The functions are defined on all mathematically possible values of the potential causes, including combinations of values that may be jointly inconsistent with the laws. For example laws $A = B$, $C = f(B,A) = B \cdot A$, are inconsistent with $A = 1$, $B = 0$, although $f(B,A)$ is defined for these values.
4. A *realization* of the system is an assignment of values to the variables. A *legal realization* is an assignment consistent with the laws.
5. A counterfactual assignment α relative to a legal realization ρ alters the values of one or more variables and determines the values of all other variables according to the laws, α, and ρ.

In the deterministic case, with binary variables, these conditions amount to assuming an acyclic graphical causal model, in which the legal dependencies of child variables on parent variables—the laws--are given by truth functions. The same formalism lurks behind various probabilistic accounts of singular causation, only differing in making each child variable a stochastic function of the values of its parents. The intervention interpretation of counterfactuals which these formal representations presume is justified, first, by the fact that interventions satisfy the Lewis axioms for counterfactuals[iii], and that almost (but not quite) all philosophical discussion of cases with explicit diagrammatic and truth functional representations make counterfactual judgements corresponding to interventions (e.g., if A causes B, on the supposition that the value of B is contrary to fact, it does not follow that the value of A is contrary to fact).[iv]

Analyses of actual causation have assumed that the relation obtains between *values* of variables representable in such networks. So, for example, we have the story of a hiker walking along a path, a boulder that rolls down the mountain above her, causing her to duck, resulting in her survival, and the question: what caused her survival? The representation as a causal model and "actual" values is then :

Boulder rolls(**true**/false) ⟶ Hiker ducks (**true**/false) ⟶ Hiker survives (**true**/false)
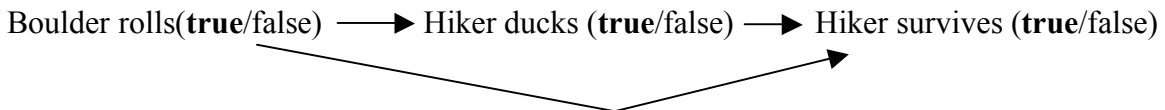
Hiker ducts = Boulder rolls; Hiker survives = ~Boulder rolls v Hiker ducks.

The actual causal relation is then supposed to be between truth values of variables, given truth values of other variables, and given a truth function for each directed edge in the graph.

Many cover stories obviously repeat the same formal structure, e.g., B throws a ball at the window S, but H catches the ball and the window survives intact. Not counting these

obvious equivalencies as distinct, the literature cited discusses about a bakers dozen examples, Our first concerns are whether this is an adequate sample, and whether an adequate sample of cases, each subjected to philosophical "intuition," is possible at all. So we will try to count how many distinct, potential examples there could be.

Restricting consideration to deterministic cases with binary variables, it would seem that even with as few as three potential causes the number of possible cases is too large for intuitions to survey. Suppose there are 3 possible causes of an effect. There are 25 possible acyclic graphical relations among the possible causes: 1 disconnected graph, 6 graphs with one edge, 6 graphs with three edges, 3 graphs with two variables having edges directed into a third (a collider), and 9 other graphs with two edges. For each one-edge graph, there are 2^2 truth functions; for each collider graph there are 2^4 truth functions; for each other two edge graph there are likewise 2^4 truth functions, and for each three edge graph on the three variables there are 2^6 possible truth functions. Altogether, then, including the disconnected graph as a trivial case, there are 601 causal models among the three potential causes.

Any subset of the three potential causes can be graphically connected with the effect variable by edges, hence there are the following possibilities: no connection, 3 single edge connections, 3 two edge connections and 1 three edge connections. For each single edge connection from possible causes to the effect variable, there are 2^2 truth functions for the effect; for each two edge connection there are 2^4 truth functions, and for each three edge, 2^8 truth functions. Altogether, then, 317 possibilities, again counting the no connection graph as one case.

Since any causal model among the potential causes can be paired with any dependency for the effect, there are 190,517 possible causal models altogether. For each of these there are $2^4 = 16$ truth value assignments, so the total number of cases for intuition to survey is 3,048,272, with just three potential causes. Intuition won't.

It may be that many of these cases can be excluded in light of general principles about the very meaning of causation principles that need no inductive justification. We will suggest one. (Hereafter, we count only pairs of graphs and truth functions—causal models-- under various restrictions. Including the variety of possible truth value assignments, the numbers below would in each case be multiplied by 2^m, where m is the number of exogenous (zero indegree) variables in the graph of a model.[v])

Every example of an actual causation in the literature that uses graphical causal models to display the laws of the system implies a further requirement connecting the (possibly stochastic) functional relations of child and parents with the graphical structure.

6. For each parent X of a variable Y, the function Y = f(Parents(Y)) allows a *test pair* for X: there are two (not necessarily legal) realizations, α and β such that for all variables Z in Parents(Y)\ X, $\alpha(Z) = \beta(Z)$, and $\beta(X) \neq \alpha(X)$ and $f(\alpha(\text{Parents}(Y))) \neq f(\beta(\text{Parents}(Y)))$

The test pair condition, (6), is logically independent of the much discussed Markov property—each variable in a DAG is independent in probability of its non-descendants conditional on values of all of its parents. The Markov condition relating DAGs and probability distributions formally allows a parent variable in a graph that is independent of its child—although such cases never occur in scientific or philosophical uses of the representation.. Given the Markov assumption, however, the test pair condition is implied by, but strictly weaker than, the Minimality condition (no proper subgraph of a graph satisfies the Markov condition for the probability distribution). So we count again.[vi]

Table 1

# parents	# truth functions	# truth functions with test pairs
1	4	2
2	16	10
3	256	218
4	65,536	64, 594
5	$> 4 \times 10^{12}$	$> 4 \times 10^{12}$

For a graph with 3 edges on three variables, one variable will have a single edge into it, with 2 possible test pair functions, and another will have 2 edges into it, with 10 possible truth pair functions, and the total number of structures meeting the test pair condition on the particular three edge, three variable graph will therefore be 20. Since there are 6 such complete directed acyclic graphs on three vertices, there will be 120 structures with three edges on three variables meeting the test pair condition. Similarly, we can count the number of causal models with smaller numbers of edges. Thus for three potential causes, and their 25 possible graphical relations, we have:

Table 2: # of truth functions satisfying the test pair condition among 3 binary variables as a function of DAG structure:

Graph forms & numbers	Number of test pair truth functions per graph	x # graphs
1 Disconnected graph	1	1
6 graphs of the form ->	2	12
6 graphs of the form -> ->	2 x 2 = 4	24
3 graphs of the form <- ->	2 x 2 = 4	12
3 graphs of the form -> <-	10	30
6 graphs of the form -> ->	2 x 10 = 20	120

Total: 199 structures.

Consider now the ways that the effect variable can depend on the three potential causal variables: it can be a function of any one of them, on two of them, or on all three of them, and the number of distinct test pair structures depends on the form of the graph, Considering only test pair truth functions we have, per graph among the possible causes
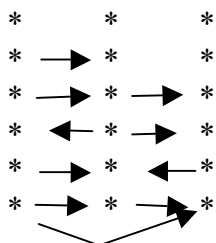
Table 3: # of truth functions of an effect variable as a function of at most 3 potential causal variables:

Graphs	#TFs, 0 args	#TFs, 1 arg	#TFs, 2 args	#TFs, 3 args	Total
(1) disconnected	1	3 x 2	3 x 10	218	255
(6) form ->	1	3 x 2	3 x 10	218	255
(6) form -> ->	1	3 x 2	3 x 10	218	255
(3) form <- ->	1	3 x 2	3 x 10	218	255
(3) form -> <-	1	3 x 2	3 x 10	218	255
(6) form -> ->	1	3 x 2	3 x 10	218	255

Since every structure among the causes is consistent with every structure between the potential causes and the effect, we have $255 \times 199 = 50,745$ structures on three potential binary causes and one binary effect. The number of causal models meeting the test pair condition with three causal variables is roughly a third of the number of graph/truth function pairs without the test pair restriction, but that is not nearly reduction enough for intuition to survey the cases.

And things get worse as the number of potential causes increases, much worse. Consider five possible causes, and *only* the case in which there are no causal connections among the causal variables, and all five causes satisfy the test pair condition for the effect, in other words the single graph in which the edges are from each of the potential causes to the potential effect, and only truth functions meeting the test pair condition for each potential cause are allowed. There are more than 4 trillion cases.

It could be argued that the number of cases is actually smaller. The idea with formal models is that structure alone counts, not the names given to variables. If we consider unlabeled directed acyclic graphs among the three potential causal variables, rather than labeled graphs, there are only the six possible structures:

```
*        *        *
*  ->  *        *
*  ->  *  ->  *
*  <-  *  ->  *
*  ->  *  <- *
*  ->  *  ->  *
```

We will simply count the number of distinct ways the effect variable can depend on three unlabeled variables. Implicitly, this also characterizes the ways in which the unlabeled variables can be truth functions of one another. The sum of the second column in table 1

now gives the number of test pair structures on three variables, 41. Table 3 is slightly altered because a truth function of X and Y is no longer distinguished from a truth function of Y and Z in structures X -> Y <- Z and X <-Y -> Z:

Table 4: Truth functions for the effect as a function of unlabeled DAGs on 3 variables

# parents of Effect:	0	1	2	3	
	#TFs	#TFs	#TFs	#TFs	Total
disconnected graph	1	2	10	218	231
->	1	3 x 2	3 x 10	218	255
-> ->	1	3 x 2	3 x 10	218	255
<- ->	1	2 x 2	2 x 10	218	243
-> <-	1	2 x 2	2 x 10	218	243
-> -> (with extra arrow)	1	3 x 2	3 x 10	218	255

The sum of the products of the entries in the 2nd column of table 2 with the corresponding entries in the total column of table r is 231 +255 x 2 + 255 x 4 + 243 x 4 + 243 x 10 + 255 x 20, or 9,956. Smaller, but still a busy time for intuitions.

We can impose further conditions, which are plausible but could conflict with some theories of actual causation.[vii] C=c cannot be an actual cause of E=e if there is no directed path from C to E. Moreover, if there is a directed path from C to E, and there is no directed path from B to E, then whether or not C=c is an actual cause of E=e cannot depend on whether or not B=b. Given this, various models are dispensable or equivalent. E.g., using the list of unlabeled graphs above, there is only one distinct graph with E as a function of a single variable and with the graph *-->*-->* among the potential causal variables, namely that in which E depends on the terminal star. The other two options merely replicate a case counted among those with *-->*   * as the relevant substructure on the causal variables. By that count, there are 20 relevantly distinct graphical structures over the three potential causes and E. When test pair truth functions for the dependencies are considered, the total number of alternative causal models 9including causal relations among the potential causes) is slightly reduced further.

Table 5: Causal models with 3 potential causes subject to restrictions

TF Multiplier For Causes	Graph	No edges to E			Total Graphs
		1	2	3	
1x	* * *	1	1	1	3
2x	*-->* *	1	2	1	4
4x	*-->*-->*	1	2	1	4
4x	*<--*-->*	0	1	1	2
10x	*-->*<--*	1	1	1	3
20x	*-->*<-*	1	2	1	4
TF Multiplier For edges to E		x2	x10	x218	

Total Models: 9682 = (1 x 1 x 2) + (1 x 1 x 10) + (1 x 1 x 218) + (2 x 1 x 2) + (2 x 2 x 10) + (2 x 1 x 218)….x (20 x 1 x 218). Smaller, but still way too big, and the number of models still grows super-exponentially with the number of variables.

What else? We can suppose that actual causation must have symmetry relations. To simplify matters, we will stop counting causal models on N variables and consider only the simpler question of the number of truth functions of N variables satisfying what may seem relevant symmetry conditions.

Say that one truth function is a *value negation* of another provided that they have the same argument variables and for each valuation of the argument variables the value of one of the functions is T if and only if the value of the other is F. Value negation is obviously a partition of the set of truth functions, for which each class has two members; it preserves the test pair relation, so the number of test pair cases is cut in half: there are 5 cases for 2 arguments, 109 for 3 arguments, but, unfortunately, more than 2 trillion for 5 arguments.

Table 6

Truth Functions for 2 arguments with Test Pair Cases

X	Y	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
1	1	1	0	1	0	1	0	0	1	1	0
0	1	1	0	1	0	0	1	1	0	0	1
1	0	1	0	0	1	1	0	1	0	0	1
0	0	0	1	1	0	1	0	1	0	1	0

The value negation partition classes are: {G1, G2}, {G3, G4}, {G5, G6}, {G7, G8}, {G9. G10}

This assumption reduces the cases by half, but two trillion cases is still far too many, as for that matter, is 5,000 cases, give or take, on 3 variables.

We might further assume that actual causation is symmetric with respect to interchange of true and false in all arguments: truth functions Gi and Gk are equivalent if Gi(X) = Gj(Xt/f) where Xt/f substitutes f for all t values in the vector X, and t for all f values. Call this relation one of *argument negation*. Equivalence under argument negation preserves the test pair property and also partitions the truth functions. The resulting classes for functions of two variables meeting the test pair condition are; {G1, G7}, {G2, G8}, {G3, G5}, {G4, G6}, {G9}, {G10}. The partition has more classes than with value negation, and the result is a smaller reduction in the search space. The numbers can be further reduced if we take the union of any two classes sharing a member, one class in each of the two partitions (i.e., classes of truth functions that are equivalent under value negation *or* argument negation). This yields: {G1, G2, G7, G8}, {G3, G4, G5, G6}, {G9. G10}. The 16 truth functions are reduced to 10 test pair truth functions, which are then reduced

to 3 classes. Combined with using the unlabeled, rather than the labeled directed graphs, the number of cases for 3 potential causes begins to seem surveyable.

The value negation partition reduces the number of functions satisfying the test pair condition by half; argument negation reduces the number of functions satisfying the test pair condition by somewhat less than half. Taking the union of the classes can reduce the number by at most a factor of 4. Unfortunately, the number of truth functions meeting the test pair condition grows super-exponentially with the number of arguments, that is, of potential causes. So for $N = 4$, there remain 16,253 cases. For $N = 5$, the reduction is to something more than a trillion cases for E as a function of 5 potential cases satisfying the test pair condition.

One further potential formal principle deserves remark. Consider the case $X \rightarrow Z \leftarrow Y$

Table 8

X	Y	Z= g1	Z= g2
T	T	F	F
F	T	T	F
T	F	F	T
F	F	T	T

Remembering that the labels of variables are not meaningful, we might argue that truth functions g1 and g2 are really the same truth function because on interchanging Y values and X values, g1 becomes g2, that is, $g1(Y, X) = g2(X,Y)$ and $g2(Y, X) = g1(Y, X)$. Each permutation of argument columns in the truth tables takes each truth function either into itself or into another truth function; The number of truth function equivalence classes that result is the original number of truth functions divided by $N!$ There are redundancies with the classes obtained from value negation and argument negation, for example, for $N = 2$, permuting arguments results in no additional reduction of classes. Nonetheless, unlike the other devices, permutation equivalence yields an exponential (as a function of N) reduction in the number of "equivalent" truth functions. But the number of test pair cases grows super-exponentially. Thus for 5 arguments, 4 trillion plus truth functions are reduced to about 35 billion classes by permutation equivalence.

We can (by computer) calculate the number of distinct equivalence classes (at least up to $N = 4$; after that, the computer takes days) of truth functions on N variables if we combine the test pair condition with value negation, argument negation, and permutation of arguments, i.e., two truth functions are equivalent if they are equivalent under any of these relations. There are then 3 classes for $N = 2$, 26 classes for $N = 3$ and 1579 classes of allowable truth functions for $N = 4$. These counts can be viewed lower bounds on the number of causal models without labels on the causal variables. Considering once again that actual causation cases as considered in the philosophical literature concern particular truth value assignments, to get a lower bound on the number of examples we must multiple these numbers by 2^{N+1}. We are back to more than fifty thousand distinct examples with four potential causes of the effect. So we consider the other recourse to

save induction by philosophical intuition from futility—that the extra cases hold nothing new. We will disprove that by examples.

4. Four Theories

For purposes of illustration, we focus on two theories in the literature that are definite enough to apply to all cases, and supplement them with two simple theories that are more less direct statements of the test pair condition plus a minimality assumption. A number of other possibilities are described in Wimberly and Glymour (in press) and in Glymour (2005).

Building off of earlier proposals by several authors, James Woodward (2003) makes the following proposal.

W

"Consider a particular directed path P from X to Y and those variables $V_1..V_n$ that are not on P. Consider next a set of values $v_1.,,,v_n$, one for each of the variables V_i. The values $vi_1…v_n$ are in what Hitchcock calls the *redundancy range* for the variables V_i with respect to the path P if, given the actual value of X, there is no intervention in setting the values of V_i to $v_1..v_n$ that will change the actual value of Y. " (Woodward, 2003; 83)

"To determine whether X =x actually causes Y = y, first apply AC.

AC:
      AC1 The actual value of X = x and the actual value of Y = y.
      AC2 There is at least one route [directed path] R from X to Y for which an intervention on X will change the value of Y, given that other direct causes Z of Y that are not on the route have been fixed at their actual values."

If AC yields an actual cause, then stop; otherwise go to AC'1 and AC'2 below.

"AC'1 The actual value of X = x and the actual value of Y = y.
AC'2 For each directed path P from X to Y, fix by interventions all direct causes Z_i of Y that do not lie along P at some combination of values within their redundancy range. Then determine whether for each path from X to Y and for each possible combination of values for the direct causes *Zi* of Y that are not on this route and that are in the redundancy range of Z_i, whether there is an intervention on X that will change that value of Y. AC'2 is satisfied if the answer to this question is "yes" for at least one route and possible combination of values within the redundancy range of the Zi. "(Woodward, 2003; 84)

There is one ambiguity in the W proposal. An intervention that changes the value of some direct cause Z_i of the value of the effect variable to some value within its redundancy range may, according to the laws of the system, also indirectly change the value of some variable on path P—indeed may change X--from its initial actual value. Is the

determination in the second sentence of AC'2 to be done with respect to the state in which the values of the variables on path P (except the effect variable) have their initial actual values, or is it to be done with respect to the values implied by the laws when the variables not on the path are fixed at their redundancy values? For most cases the ambiguity does not matter, but we will note a case in which it introduces complexities.

Joseph Halpern and Judea Pearl have recently made a different proposal:

HP2005

"$(M, \mathbf{u}) \models [X <\text{-} x] \varphi$" abbreviates '$\varphi$ is true in structure M for legal realization $\mathbf{u}$ if $\mathbf{u}$ is possibly altered by an intervention setting X to value x.' Boldface denotes sets, variables are uppercase and their values the corresponding lower case. Thus $\mathbf{X}$ is a set of nodes or variables, $\mathbf{x}$ a set of values, and $\mathbf{X} = \mathbf{x}$ denotes that the variables in $\mathbf{X}$ have values $\mathbf{x}$.

$\mathbf{X} = \mathbf{x}$ is an actual cause of φ in $(M, \mathbf{u})$ if and only if:

AC1; $(M, \mathbf{u}) \models (\mathbf{X} = \mathbf{x})$ & φ

AC2: There exists a partition $(\mathbf{Z}, \mathbf{W})$ of $\mathbf{V}$ with $\mathbf{X} \subseteq \mathbf{Z}$ and some setting $(\mathbf{x}', \mathbf{w}')$ of the variables in $(\mathbf{X}, \mathbf{W})$ such that if $(M, u) \models Z = z\bullet$ for all $Z \in \mathbf{Z}$ then both of the following conditions hold:

(a) $(M, \mathbf{u}) \models [\mathbf{X} <\text{-} \mathbf{x}', \mathbf{W} <\text{-} \mathbf{w}']\sim\varphi$

(b) $(M, \mathbf{u}) \models [\mathbf{X} <\text{-} \mathbf{x}. \mathbf{W}' <\text{-} \mathbf{w}', \mathbf{Z}' <\text{-} \mathbf{z}\bullet]\varphi$

for all subsets $\mathbf{W}'$ of $\mathbf{W}$ and for all subsets $\mathbf{Z}'$ of $\mathbf{Z}$. In words, setting any subset of variables in $\mathbf{W}$ to their values in $\mathbf{w}'$ should have no effect on φ, as long as $\mathbf{X}$ is kept at its current value $\mathbf{x}$, even if all the variables in an arbitrary subset of $\mathbf{Z}$ are set to their original values in the context $\mathbf{u}$.[viii]

AC3: X is minimal; no [proper] subset of $\mathbf{X}$ satisfies conditions AC1 and AC2.

AC4: X = x & $\sim\varphi$ is consistent

Any actual cause according to AC of W is an actual cause according to HP 2005. We will show by example that not every W actual cause is an HP 2005 actual cause.

Each of the following stories, variants of which are all over the literature, corresponds to a set of truth functional relations among propositional variables and a valuation of the variables. The truth functional relations and valuations can in all cases be realized with switches and lights in electrical circuits, or with computer logic chips. Indeed, they are realizable in any computer. Whatever ambiguities of background and context may attach to the stories are removed in the physical models, and an adequate theory of actual causation must, therefore, account for causation in the truth functional systems.

1. A and B each fire a bullet at a target, simultaneously striking the bullseye (D). What caused the bullseye to be defaced?
 $A \to D \gets B \quad D = A + B; A = B = D = 1$

W, HP2005: Actual causes of D =1 are A = 1 and B = 1

2. A and B each fire a bullet at a target. A's bullet travels faster, knocking out the bullseye (D), which B's bullet would have knocked out a moment later (D') otherwise. What caused the event D = 1, of the bullseye's removal?
    B -> D' <- A -> D; D = A, D' = B(1 – A); A = B = D = 1; D' = 0

   W, HP 2005: The actual cause of D = 1 is A = 1.

3. A and B each fire which would have missed the target, except that the bullets collide (C = 1) and A's bullet ricochets into through bullseye. What caused the bullseye to be hit (D = 1)?

   A ⟶ C ← B   C = A•B; D = C; A = B = C = D = 1.
         ↓
         D
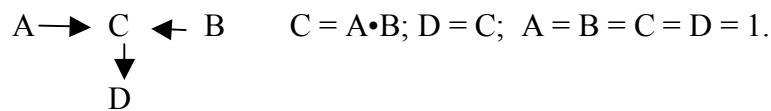
   W, HP 2005: The actual causes of D = 1 are A = 1, B = 1, and C = 1.

4. A, a perfect marksman, is about to fire at the bullseye; B is about to jostle A to prevent A from hitting the bullseye; C shoves B out of the way. A fires and hits the bullseye (D). What caused the bullseye to be hit?
    C -> B -> A -> D; D = A; A = (1 – B); B = (1 – C); A = D = C= 1, B = 0
   W, HP2005 : The actual causes of D = 1 are A =1, C = 1, and B = 0

5. A, an imperfect marksman, is about to fire at the target, but his aim is too low. B standing at the back of the crowd, could push his way through to A and lift the rifle barrel just the right amount, but B does no such thing. A's bullet misses the bullseye. What caused the bullseye to be missed (D = 0)?
    B ⟶ A ⟶ D   D = (1 – A), A = (1 – B) B = 0, A = 1, D = 0.
   W, HP 2005: The actual causes of D = 0 are B = 0 and A = 1

6. A, the perfect marksman, aims (A) at the target, but (C) fails to cock his gun, and (P) pulls the trigger. The gun does not fire and the target is untouched. Which event caused the gun not to fire (F = 0)?
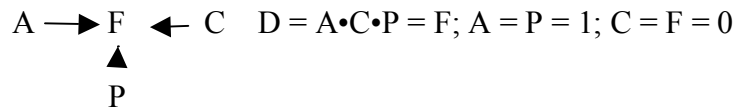
   A ⟶ F ← C   D = A•C•P = F; A = P = 1; C = F = 0
         ▲
         P
   W, HP 2005: The actual cause of D = 0 is C = 0

7. As gun has a safety mechanism: the gun will not fire unless the hammer is cocked and the round is chambered and the trigger is pulled. Pulling the trigger causes a round to be chambered but prevents the hammer from being cocked. The trigger is pulled. The gun does not fire. What caused the gun not to fire (F = 0)?

$H = (1- T), C = T, F = H•C; T = C = 1, H = F = 0.$
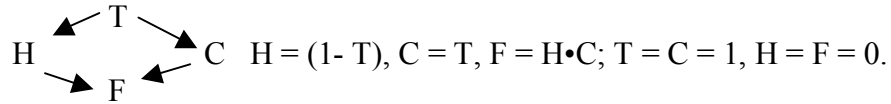
W, HP 2005: The actual cause of $F = 0$ is $H = 0$.

8. The right hand of the ambidextrous perfect marksman is bitten by a dog; he pulls the trigger with his left hand and hits the bullseye. What caused the marksman to pull the trigger with his left hand? What caused the bullseye to be hit?

$B \to H \to D$ $H = 2$ if $B = 1$; $H = 1$ otherwise; $D = 1$ if $H = 1$ or 2; $D = 0$ if $H = 0$. $B = 1, H = 2,, D = 1$

W, HP 2005: $B = 1$ caused the left-handed shot ($H = 2$)
W, HP 2005: $H = 2$ caused the bullseye hit ($D = 1$); $B = 1$ did not cause it

9. A boulder slides ($B =1$) toward a hiker, who, seeing it, ducks ($D = 1$). The boulder misses him and he survives ($S = 1$). Did the boulder sliding cause his survival?
$B \to D \to S$ $S = (1- B) + D ; S = B = 1$

W, HP 2005: The actual cause of $S = 1$ is $D = 1$.

10. A and B, both perfect marksman, shoot at the target at almost the same time. The ejected shell from A's pistol deflects B's bullet ($C = 1$), which would otherwise have hit the target bullseye. A's bullet hits the bullseye. What caused the bullseye to be hit ($H = 1$).

$C = B • A, H = B • (1- C) + A. A = B = H = C = 1$
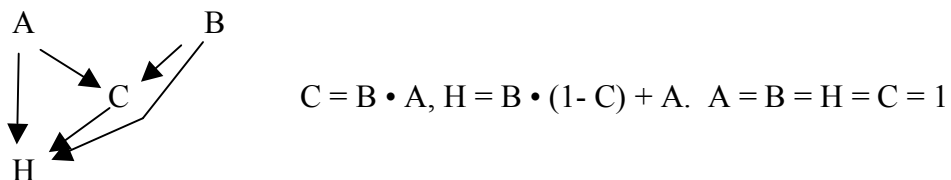
W: The actual causes of $H = 1$ are $A =1, B = 1, C = 1.$ [ix]
HP: the actual cause is $A = 1$

11. A and B, both perfect marksmen, pull their triggers on similar guns at the same time. B loaded her rifle ($Lb = 1$) and hits the bullseye ($H = 1$). A has forgotten to load his rifle ($La = 0$). What caused the hit?
$H = A•La + B•Lb$ $A = B = Lb = H = 1; La = 0$

W, HP 2005: The actual causes are $B =1$ and $Lb = 1$

12. A, B, C, D and E fire at and simultaneously hit a target that will fall over if at least 3 bullets hit it. The target falls over (F =1). What caused the target to fall over?

W, HP 2005: The actual causes of F = 1 are A =1, B =1, C =1, D = 1, and E = 1.

13, A woman takes birth control pills (B) which prevent a pregnancy (P) that, had it occurred, would have caused the thrombosis (T) caused by taking the birth control pills.

$$B \longrightarrow P \longrightarrow T$$
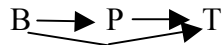
B -> P -> T   P = (1 – B); T = B + P. B =T = 1, P =0

W, HP2005: The actual cause of T = 1 is B =1

14. A and B have three mutually exclusive choices, to vote for C, or for D, or not to vote . An option wins if A votes for it or if B votes for it and A does not vote (= N). A and B vote for C (A, B = c).

W; A = c is the actual cause of C winning .
HP 2005: A =c, B = care the actual causes of C.

There are any number of proposals that agree with many of the judgments of both W and HP 2005 on these cases. We give two simple proposals—neither of which we endorse-- that disagree with W and HP only slightly in these cases.

Simple:

The actual value x of a variable X is an actual cause of the actual value y of a variable Y in a state s of a system if and only if there is a value y' ≠ y for Y, and X is a member of a set **X** of variables (not having Y as a member, of course) with actual values **x,** and there exist alternative values **x',** none of which equal the corresponding values in **x,** such that an intervention on the system in state s that fixes **X** = **x'** entails Y = y', and no proper subset **Z** of **X** with actual values **z** is such that there exist alternative values **z',** none of which equal the corresponding values in **z,** such that an intervention on the system in state s that fixes **Z**= **z'** entails Y = y'.

SimpleJ: replace 'no proper subset' in Simple by "no set of lower cardinality."

For case 10, the Simple theories both say that the actual causes of H = 1 are the members of the set A = 1, B = 1]--change A and B to 0 by intervention, and C changes to 0 and H changes to 0, but H is not 0 if either A or B alone changes to 0, . For case 13, the Simple theories both say that there are *no* actual causes of thrombosis. For case 14, the Simple theories both agree with W.

14

5. Looking Further

The enormous space of alternative causal structures would be of little interest to the discussion if it contained no puzzling cases that raise issues not already present with three or fewer potential causes, or that reshuffle the alliances among proposed analyses. We will show that there are such cases. How many we cannot say, and that is the point.

Consider the example:

$$A \rightarrow B \rightarrow D \rightarrow E$$

with C above, arrows from C to B, D, and E.
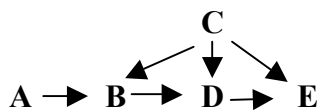
$E = C + D$; $D = B * C$; $B = A * C$. $A = B = C = D = E = 1$.

Which of these are actual causes of $E = 1$?

W; The actual cause is $C = 1$.
HP 2004, The actual causes are $C = 1$, $D = 1$

Simple, SimpleJ; The actual cause is $C = 1$.

People may differ at first consideration as to which answer is correct. Indeed, someone who holds that causality is transitive might think that none of the answers is correct, arguing that if $D = 1$ is a cause of $E = 1$, then so must $B = 1$ and $A = 1$ be, since $D = 1$ is caused by $B = 1$ and $B = 1$ is caused by $A = 1$.

Consider another more interesting case, this time with five variables. On a ranch there is a complicated rule: everyone, including the Cowboy and the Ranger, vote on what is to be done; the Ranger's vote is the outcome if the Cowboy and the Ranger agree, or if the Ranger Stands Alone and everyone disagrees with her; otherwise, majority rules. The Cowboy and the Ranger vote for a round-up; the two Hands and the Wrangler vote to stay around the campfire. Did the Wrangler's vote cause the round-up?

We need first to consider whether voting against a round-up is strategic: sometimes a vote superficially *against* is really a vote *for*. The ranch is not such a case. One sense of what a vote is *for* is what it would rationally be if a specific outcome were desired. Assume Wrangler wanted not to go on a round-up and Wrangler is in ignorance about how all of the others will vote: his priors for every vote but his are 50/50 for round-up. No matter how Wrangler votes, cases in which Cowboy and Ranger agree on 0 (= stay by the campfire) are equally likely as cases in which Cowboy and Ranger agree on 1 (= go on a round-up). Averaged over these cases, Wrangler is as likely to get his desire if he votes 1 as if he votes 0. Ignore them. That leaves $2^3 = 8$ equally likely voting patterns for

15

others than Wrangler. In 2 of these patterns Ranger stands alone, and Wrangler has averaged over these cases an equal chance of getting his desire if he votes 0 as if he votes 1. Ignore them. There remain 6 cases in which Ranger and Cowboy do not agree and the Ranger does not stand alone (ignoring Wrangler's as yet undecided vote). They are:

Cowboy	Ranger	Wrangler	Hand 1	Hand 2	Wrangler/Round-up	
1	0	?	1	0	0/0,	1/1
1	0	?	0	1	0/0	1/1
1	0	?	0	0	0/0	1/0
0	1	?	1	0	0/0	1/1
0	1	?	0	1	0/0	1/1
0	1	?	1	1	0/1	1/1

If Wrangler votes 0, then Round-up = 0 in the cases in the first 5 rows. If Wrangler votes 1, then Round-up = 0 in the 3rd row only. In voting 0, Wrangler is voting against a round-up, and so are the Hands.

It remains to say which votes are causes of going on the round-up under which proposals. The answers are;

W, SimpleJ: R =1^x
HP, Simple: R = 1; Wrangler = 0 Hand = 0, Hand 2 = 0

Things come apart in a novel way in this case—just one Hand won't create the same puzzle, and the answer HP gives arises for a surprising reason. [xi] What perplexities lurk elsewhere among the billions of unexamined examples?

6. Misrepresentation and Metaphysics

On occasion, Bertrand Russell mocked traditional philosophers for creating paradoxes by treating a relation as a monadic property and equivocating over one of the relata. We suggest that something of the same kind is at work in much of the philosophical literature on actual causation.

The Bayes net representation of causal systems (Spirtes, et al., 1993) was developed for representing causal relations among systems of variables as they are deployed in engineering, medicine and the natural and social sciences, including a characterization, given a causal model, of the effects of exogenous *changes* in any set of variables on any disjoint set of variables. No notion of the value of one set of variables causing the value of another variable was part of that characterization, titled the Manipulation Theorem. The adaptation of Bayes nets for descriptions of actual causal relations attempts to introduce a causal relation between values of variables in a single state of the system given a causal model. Alternative states are only referenced in the counterfactual or intervention conditions of the analyses. As Halpern and Pearl note. their actual causes are not changes, but possible worlds: "Note that we are using the word 'event' here in the standard sense of 'set of possible worlds' (as opposed to 'transition between states of

affairs'); essentially we are identifying events with propositions." (2005, n. 6). And that is part of the problem.

We tend to think of causes as changes, or happenings. On one reading of the diagrams, a distinguished value (say "1") represents the occurrence of some event, and the other value (e.g., "0") represents the *absence* of that event. The event itself might or might not be a change of some feature or condition of some space time-region. The absence of the event is often nothing definite at all, which is one source of worry about the vagueness of counterfactuals and about the actual causal relevance or irrelevance of absences. In some cases, the imposition of a Bayes net representation on a causal story about events forces a false disambiguation of both presences and absences, as though there were always laws constraining relations between occurrences or absences of some events and occurrences or absences of previous events.[xii]

An alternative interpretation of the diagrams was already suggested by Lewis' description of marked directed graphs as "neuron diagrams"—systems with spatially localized parts that have distinguishable *states*. The same is true of wiring diagrams. The states of the system in this case do *not* represent changes or happenings, and in most cases the problem of what caused what, given a causal model and its state, is underspecified. We can intelligibly ask what caused the state of a variable, or what caused a change in its state, but relevant answers are typically—conceivably, there are exceptions—about what other changes of other variables brought about that state (from another, or none) or change of state. For a neuron diagram, actual causal relations typically involve at least two total states, each specifying values for the local variables of the system, and it is the change over time of values for some variables between the two states that brings about a change in others, or prevents them from changing when they otherwise would have from other changes. In a single diagram with a single value given for each vertex, intuitions about what causes what may vary because, implicitly, people make different assumptions about the prior states. Outside of formal representations, this gets parsed as "normal conditions' or "the causal field" or perhaps "defaults" and in discussions of actual causation is generally left inexplicit. Informality is not a solution to equivocation.

Consider the possible transitions from left to right in the state of a system, where we use Lewis' convention that A-> B o- C means that B = A(1 − C), with A, B, C taking values in {0,1}. Dark vertices code 1 while empty vertices code 0.

1.

2.

3.

4.
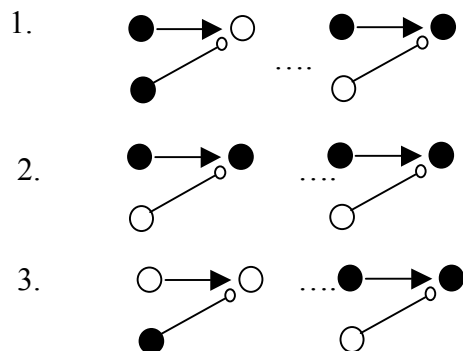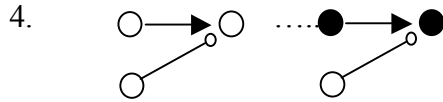
In all four cases, the final state is the same, but we wager that many people, shown the sequences—or their equivalents in some less abstract representation of the same structures and state relations--would not judge the causes of the final state of the right-most node to be the same in all four cases.[xiii] We expect that common judgements would locate the cause in three of these sequences to be the changes of states of another node or nodes Whether in the second sequence *anything* would be commonly judged to be the cause of the final state of the right-most node seems an interesting question.

The changes of values of nodes can of course themselves be represented as nodes, with a different interpretation for each of the above sequences. For example, the *changes* in the third sequence above:
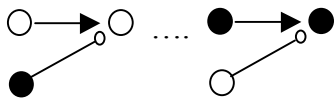

might be represented as
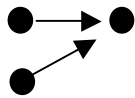
where a dark node indicates that an event occurred—a local change of state. But the change (or happening) graph representation has no clear functional dependencies that are independent of the actual beginning and end states—no laws—and fails to mark the difference between a change in a node from empty to dark, and a change of that same node from dark to empty—each kind of change becomes a dark node. Thus the transition
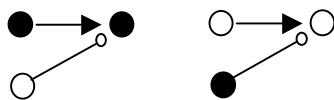

would be represented by the same "change graph" above.[xiv] There are other ambiguities when actual causes are not understood as changes.

In Bayes nets any old process can be inserted between two variables related by a directed edge: A -> B becomes A -> pretty much anything you want -> B. The probability relations, intervention relations, and variable causation between A and B can remain unaltered, but arguably in some cases the actual causation relations are changed. Ned Hall (2004) has pointed out that the diagram and truth function:


$E = A \vee B$
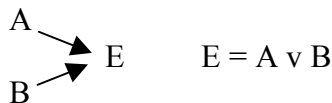
is consistent with the mechanism:

A
C → E   E = C v A; C = B & ~A
B
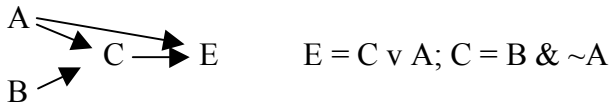
Evidently, when A= B = E = 1, B has no role in producing E.   According to all of the proposals considered, A= 1 and B = 1 are actual causes of E = 1, with or without the intervening mechanism, and Hall suggests the same. What would others say?

All of these problems vanish if we consider changes that changes in a particular system state produce. The Manipulation Theorem then gives a relation between a system state, changes that are exogenous ideal interventions—the very interventions traded on in the counterfactuals of the counterfactual analyses of actual causation—and changes in other variables. The theorem is a necessary consequence of a fundamental principle about causal models, the Markov Property, which assumed in all of the discussions we have mentioned. Strengthening the Markov assumption with Mimimality—the test pair condition--then permits an algorithm for computing the changes an ideal intervention produces (Pearl, 2000). No induction over cases is required. Various problems disappear, Hall's for example: given the state of the system, and given an intervention on a variable (or variables), the resulting changes in the state of other variables is unambiguous. Transitivity for another: if a particular change produces another change, and *that* still another, then the first produces the third. Disputes over actual causation look like the fruit of misleading representations.

Which is not to say that no puzzles remain. Causal explanations of non-changes by non-changes are common enough:: *the rains did not flood the valley because the dam held.* Or consider the following example (due to Peter Spirtes): A sets a dial at time t; B sends a message at t to C as to the final dial setting wanted; C examines the dial setting at t'> t, and turns the dial the smallest distance so that the result accords with B's desire. The result is a dial setting at t''> t'. The question is: what caused the setting of the dial at t''? The Bayes net graph looks something like this:
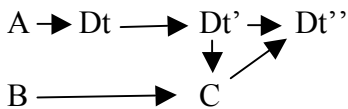
A → Dt ⟶ Dt' → Dt''
B ⟶ C

Suppose that the setting A gives to the dial, Dt = Dt', is the setting that B desires. Then C does nothing to the dial setting—never touches it—and Dt = Dt' = Dt''. To several intuitions, A's action is the cause of the value of Dt'', but there is no counterfactual dependence of Dt'' on A of the sort: *If A had given the dial a different value then Dt'' would have had a different value.* There is only the double counterfactual: if A had given the dial a different value *and C had still done nothing*, then Dt'' would have had a different value. It is difficult to see how to analyze such cases in terms of the values of individual nodes, because it is a *relation* between values that underlies the anomaly, if that is what it is. The effect of equal values received at C and Dt' is *as if* to remove the

edge between C and Dt''. In this case, W, Simple and SimpleJ find the values of B and of C to be the actual causes of the value of Dt'', which seems wrong. HP 2005 says the values of B, of C, of A, of Dt and Dt' are all causes of the value of Dt'', which seems wrong again.

7. Whose Judgement?

The presumption that philosophers' judgements in puzzling cases are or ought to be authoritative is at once comforting and unwarranted. There is no reason why the issues in particular cases cannot be explained to a wide range of people, and their responses explored. One would like to know the distribution of informed opinions that would reject as ambiguous the very question of actual causation given one or another description of circumstances. One would like to know the distribution of judgements in cases like Hall's, both before and after the mechanism is revealed. One would like to know how people judge cases in which the causes are apparently ineffectual. One would like to know how people judge cases that are structurally like various voting arrangements, with and without a cover story that suggests voting. One would like to know whether judgements of causation for localized variables depend only on the final state or on the transitions that lead to it. One would like to know when and in what respects systems become too complex for people to give more than random judgements, or none at all. And much more.

Given the legal and other import, one would have thought psychologists would be all over the topic of actual causation, and we would know the distribution of adult judgements about these issues and many more. (Given the increasing number of "laboratories" in Philosophy departments, one might even have hoped some philosophers would have addressed the questions). Not so.

There is an enormous psychological literature on human judgement about causation when the joint occurrences of features are repeated (i.e., about type-level causation), and about token causation for extremely simple "mechanical" cases (e.g., collisions of objects, inspired by Michotte, 1954). A study by Sloman and Lagnado (2002) argues that in causal contexts people do not backtrack on counterfactuals. There is also some work on token causal judgement imbedded in morally fraught contexts (e.g., Ahn & Kalish, 2000; Ahn, Kalish, Medin, & Gelman, 1995; Wolff & Song, 2003), and in social contexts. In particular, Choi, Nisbett, and Norenzayan (1999) focus on causal attributions in a variety of social situations by participants in a range of cultures. Their major conclusion is that participants in Asian cultures are more inclined towards situationism: they are more likely to attribute people's actions to situations, rather than dispositional or personality traits of the individual.
There is a much more limited psychological literature on the kinds of cases philosophers have considered in discussions of token causation. Perhaps the most relevant piece of psychological work is Walsh and Sloman (2005). They provided experimental participants with a range of "standard problems" from the philosophy literature, including overdetermination, late preemption, and interruptions (A is going to cause E but B intervenes by blocking B; did A cause E not to happen? Did B cause E not to happen?).

Their results were decidedly ambiguous: except in the clearest cases—the ones on which the entire philosophical community agrees—the modal description of a situation was provided by 60% or fewer of the participants. Naïve intuitions were, for their study, no more settled than those of the philosophical community. There was one clear finding in their study: 'prevent X' was not the equivalent to 'cause not-X' for their participants. Depending on the exact story, participants would sometimes think that one or the other of these two constructions was appropriate, but they very rarely found them to be interchangeable. The experiments in Walsh and Sloman (2005) focus on a very limited domain: all of their stories use people as the potential causes, and various physical events as effects (e.g., a coin falling on heads). As they note, there is no particular justification for thinking that their results would hold if the effect were an event involving another intentional agent, or if the claims involved social causation, or if the potential causes were *not* intentional agents.

7. Conclusion

Causal Bayes nets developed as a formalism for representing causal relations among variables and for studying inferences to such relations and their use in predicting the effects of interventions. That framework is now used more or less without comment in several areas of science. It was natural enough then to take Bayes nets as a framework for actual causation, but it is a mistake to take actual causation to generally be a relation among values of nodes in such a structure, just as it is a mistake to induce vast generalizations about conditions for causal attribution from a baker's dozen of examples.

Our argument is not for an abandonment of formal representations of actual causation, or for promulgating more examples without formal control, or even for abandoning neuron diagrams or Bayes nets or graphical causal models in philosophical investigations of causal relations. It is an argument against the adequacy of the inductive method that has dominated philosophical discussion of actual causation, against the sufficiency of Bayes net representations without consideration of state transitions, and against the presumption that, in judging cases, philosophers know best

References

Ahn, W., & Kalish, C. W. (2000). The role of mechanism beliefs in causal reasoning. In F. C. Keil & R. A. Wilson (Eds.), *Explanation and cognition* (pp. 199-225). Cambridge, MA: The MIT Press.

Ahn, W.-K., Kalish, C. W., Medin, D. L., & Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition, 54*, 299-352.

Choi, I., Nisbett, R. E., & Norenzayan, A. (1999). Causal attribution across cultures: Variation and universality. *Psychological Review, 125*, 47-63.

Gilles, D. (2005). An Action-Related Theory of Causality, The British Journal for the Philosophy of Science, 56: 823-842.

Glymour, C. Review of "Causality and Chance" edited by D. Dowe and P. Noordhoff. *Mind*, 2006.

Glymour, C. and F. Wimberly, "Actual Causation and Thought Experiments" in press.

Hall, N. (2004). Two Concepts of Causation. In J. Collins, N. Hall and L. Paul, eds. *Causation and Counterfactuals*. Cambridge, MA: MIT Press.

Halpern, J and J. Pearl, (2005). Causes and Explanations: a Structural-Model Approach. Part I: Causes. *British Journal for the Philosophy of Science*, 56, 853-887.

Halpern, J. and J. Pearl (2000). *Causes and Explanations: A Structural Model Approach*. Technical report R-266. Cognitive Systems Laboratory. University of California at Los Angeles.

Hitchcock, C. (2001) The Intransitivity of Causation Revealed in Equations and Graphs" *Journal of Philosophy* 98, 273-99.

Kvart, I. (2004a). Probabilistic Cause, Edge Conditions, Late Preemption and Discrete Cases. In P. Dowe and P. Noordhof, *Cause and Chance*. New York, Routledge

Kvart, I. (2004) Causation: Probabilistic and Counterfactual Analyses. In J. Collins, N. Hall and L. Paul, eds. *Causation and Counterfactuals*. Cambridge, MA: MIT Press.

Lewis, D. (1986). Causation. In *Philosophical Papers*, Volume II, New York: Oxford University Press.

Lewis, D. (2004) Causal Influence. In J. Collins, N. Hall and L. Paul, eds. *Causation and Counterfactuals*. Cambridge, MA: MIT Press

McDermott, M. (1995) Redundant Causation. British Journal for the Philosophy of Science, 46, 523-544.

Mackie, J. (1974) *The Cement of the Universe*. New York: Oxford University Press.

Menzies, P. (2004). Difference Making in Context. In J. Collins, N. Hall and L. Paul, eds. *Causation and Counterfactuals*. Cambridge, MA: MIT Press.

Michotte, A. (1954) La **perception** de la causalité. Publications Universitaires de Louvain

Noordhof, P. (2004) Prospects for a Counterfactual Theory of Causation. In P. Dowe and P. Noordhof, *Cause and Chance*. New York, Routledge.

Novick, L.R., & Cheng, P.W. (2004). Assessing interactive causal influence. *Psychological Review, 111*, 455-485.

Nute, D. (1976) David Lewis and the Analysis of Counterfactuals." Noûs 10:455-461

Pearl, J. *Causality* (2000). New York: Oxford University Press.

Paul, L. (2004) Aspect Causation. . In J. Collins, N. Hall and L. Paul, eds. *Causation and Counterfactuals*. Cambridge, MA: MIT Press.

Ramachandran, M. (2004) Indeterministic Causation and Varieties of Chance Raising. In P. Dowe and P. Noordhof, *Cause and Chance*. New York, Routledge.

Ramachandran, M. (2004a) A Counterfactual Analysis of Indeterministic Causation. In J. Collins, N. Hall and L. Paul, eds. *Causation and Counterfactuals*. Cambridge, MA: MIT Press.

Schaffer, J. (2000). Trumping Preemption. *Journal of Philosophy* XCVII, 165-81.

Scheines, R. and P. Spirtes (2004), Causal Inference of Ambiguous Manipulations" *Philosophy of Science*, 71, 833-845.

Sloman, S.A., & Lagnado, D. (2002). Counterfactual undoing in deterministic causal reasoning. Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society, Maryland.

Spirtes, P., C. Glymour and R. Scheines, (1993). *Causation, Prediction and Search*, New York, Springer.

Spohn, W. (2005) Causation: An Alternative. The British Journal for the Philosophy of Science. 57: 93-119.

McKenzie, C. R. M., & Nelson, J. D. (2003). What a speaker's choice of frame reveals: Reference points, frame selection, and framing effects. *Psychonomic Bulletin and Review, 10*, 596-602.

Sher, S., & McKenzie, C. R. M. (2006). Information leakage from logically equivalent frames. *Cognition, 101*, 467-494.

Walsh, C. R., & Sloman, S. A. (2005). The meaning of cause and prevent: The role of causal mechanism. In B.G. Bara, L. Barsalou & M. Bucciarelli (Eds.), *Proceedings of the 27th annual conference of the cognitive science society* (pp. 2331-2336). Mahwah, NJ: Lawrence Erlbaum Associates.

Wolff, P., & Song, G. (2003). Models of causation and the semantics of causal verbs. *Cognitive Psychology, 47*, 276-332.

---

[i] 1 = Department of Philosophy, Carnegie Mellon University; 2 = Florida Institute for Human and Machine Cognition; 3 = Department of Philosophy, Kansas State University; 4 = School of Humanities, California Institute of Technology.

[ii] See, for example: Lewis (1980), Hitchcock (2001), Woodward (2003) and Halpern and Pearl (2000; 2005), among many others.

[iii] Lewis' axioms do not imply that A V B -> C |= A -> C V B -> C, which Pearl (2000) suggests is required by interventions (and Nute (1976) thinks is required by counterfactuals).

[iv] For example, Lewis, (1986), Hitchcock, (2001), Woodward, (2003), Hall (2004), Ramachandran (2004a, 2004b), Kvart (2004a, 2004b), Noordhof (2004), and Halpern and Pearl, (2005a, 2005b). Menzies (2004) describes qualitatively all of the elements of the representation without mentioning it. The stochastic version of the framework is essentially a Bayes net, with distributions satisfying the causal Markov condition (Spirtes, et al.,, 1993). Kvart's conditions, for example, are finely constructed to take advantage of the constraints the causal Markov condition imposes on relations between probability and an acyclic binary relation representing causes, but he does not specify the Markov constraint explicitly. The construal of the antecedent in counterfactuals as an intervention result is not always consistent in these papers. Hall, for example, seems to need an intervention account (which prevents backtracking) for some of his arguments (Hall, 2004; 261-262) but writes in terms of more general counterfactuals that allow them.

[v] Ternary variables have played a role in discussions, but we can make our point without considering them. Adding a ternary cause considerably increases the counts.

[vi] A closed form counting formula for truth functions satisfying the test pair condition is the continuation for each n up to n = N of $2^{2N} - n(2^{2(n-1)} + 2^{2(n-2)} + (n\text{-}1 C\ 1)\ 2^{2(n-2)} - (n\text{-}1 C\ 2)\ 2^{2(n-3)}$..... The recursive form is: F(0) = 2, and $F(n) = 2^{2^n} - \sum_{i=1}^{n} \binom{n}{n-i} F(n-i)$.

[vii] For example, that C is a cause of E if and only if C and E occur and the probability of E is higher given C than given the absence of C.

[viii]. For unexplained reasons, Halpern and Pearl restrict the scope of their definition to variables that have positive indegree, or in econometric terms, are endogenous. Any causal model can be expanded by adding, for each exogenous variables, a new variable with zero indegree and unit outdegree, directed into the originally exogenous variable, with becomes endogenous, with the variable values related by the identity function. We will therefore ignore the restriction in what follows, as do they in discussing examples.

[ix] This case is not entirely clear. The first clause of W, AC, does not apply. The redundancy range of B, C for A = 1 and for the path consisting of the directed edge from A to the effect variable is {0,1} X {0,1}; and the redundancy range of A for B = 1 and the B -> C -> H path from B to H is {1, 0}; for the B -> H path the redundancy range for B = 1 of A, C is <1, 1>, <1, 0>, <0, 0>; the redundancy range of A, B, for C = 1 is <1, 0>, <1, 1>, <0, 1>, the last since setting the value of A to 0 would change the value of C to 0 and leave B at 1, which would leave the value of the effect, H, unchanged at 1.

Changing A from 1 to 0 with C = 0, B = 0 changes the effect value, so A = 1 is an actual cause. For initial actual value B = 1, fixing the value of A at 0 but leaving all variables on the path B -> C at their actual values (B = 1, C = 1) and keeping all other implied laws (i.e., the dependency of C on B for fixed A and the dependency of H on C for fixed A) yields H is 1 if and only if the value of B is 0. If instead, fixing A at 0 implies that, with actual value B = 1, the value of C to be used in AC'2 for the B -> C > –> H path should be 0, then H is 1 if and only if B is 1. In either interpretation, B = 1 is an actual cause, although tin the first interpretation, in a very odd way.   For the path from C to the effect variable, if A is set at 0, C changes to 0, and the effect H remains unchanged at 1; if now, while leaving B at 1, an intervention changes C back to its original actual value, C = 1, while leaving A = 0, the effect variable changes value to 0. So we have the oddity that an event, C = 1, whose only apparent role is to prevent a causal sequence that would otherwise have led to the effect, is counted as an actual cause of the effect. These may not be the consequences Woodward intended, but they are what the wording implies.

For HP, X ={A}, Z = {A, B, C} and W = empty set witnesses that A = 1 is an actual cause of H = 1. For X = {B,}, W = {A}, Z = {B,C}, the actual setting entails H = 1; setting A = 0 and B = 0 entails B = 0 and H = 0, but changing B back to 1 while leaving A and C at 0 does not change H back to 0, hence this does not witness that B = 1 is an actual cause, and this same argument works for W ={A, C}. is the only possible witness for that proposition. So B = 1 is not an actual cause of H = 1. For X = {C}, W = {A}. Z = {C, B}, setting A = 0 and leaving C = 1 changes H to 0. But leaving A = 0, and changing C back to its original value, i.e., leaving it at 1, does not change H back to 0. B is irrelevant in any case. So C = 1 is not an actual cause of H = 1. Neither, it turns out, is 'B and C'.

[x] On the fourth reading of redundancy.
[xi] In HP2005, let X = Wrangler, and let W = Cowboy. Change Cowboy to 0 and Wrangler to 1. Then the Ranger does not stand alone, and majority rules, so the Roundup = 0. Now change Wrangler back to 0, leaving Cowboy at 0. Now the Ranger stands alone, so Roundup = 1. Returning Wrangler to his original state thus brings about the original result, but in a different way.
[xii]*If Napoleon had not been born, he would not have been defeated at Waterloo,* is a true counterfactual. *Napoleon's non-birth* is a metaphysical contrary of an actual event, Napoleon's actual birth, but there are a great many possible events of which Napoleon's non-birth is the metaphysical contrary.
[xiii] Exactly this type of description dependence on prior state has been found in various non-causal settings, such as descriptions of water level in a glass (e.g., McKenzie & Nelson, 2003; Sher & McKenzie, 2006).
[xiv] The literature may sometimes have an implicit reading of the graphs—it is hard to say—in which a dark node is understood to have changed and a white node is understood not to have changed.