Synthese
DOI 10.1007/s11229-009-9497-9

Actual causation: a stone soup essay

Clark Glymour · David Danks · Bruce Glymour · Frederick Eberhardt · Joseph Ramsey · Richard Scheines · Peter Spirtes · Choh Man Teng · Jiji Zhang

Received: 31 March 2008 / Accepted: 1 March 2009

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) cases with larger numbers of causes generate novel puzzles; (3) "neuron" and causal Bayes net diagrams are, as deployed in discussions of actual causation, almost always ambiguous; (4) actual causation is (intuitively) relative to an initial system state since state changes are relevant, but most current accounts ignore state changes through time; (5) more generally, there is no reason to think that philosophical judgements about these sorts of cases are normative; but (6) there is a dearth of relevant psychological research that bears on whether various philosophical accounts are descriptive. Our skepticism is not directed towards the possibility of a

C. Glymour (✉) · D. Danks
Carnegie Mellon University and Florida Institute for Human and Machine Cognition, Pittsburgh, PA, USA
e-mail: cg09@andrew.cmu.edu

B. Glymour
Kansas State University, Manhattan, KS, USA

F. Eberhardt
Washington University in St. Louis, St. Louis, MO, USA

J. Ramsey · R. Scheines · P. Spirtes
Carnegie Mellon University, Pittsburgh, PA, USA

C. M. Teng
Florida Institute for Human and Machine Cognition, Pensacola, FL, USA

J. Zhang
Lingnan University, Tuen Mun, Hong Kong

Published online: 20 March 2009

correct account of actual causation; rather, we argue that standard methods will not lead to such an account. A different approach is required.

Keywords Actual causation · Bayesian networks · Combinatorics · Intervention · Intuitions

> Once upon a time a hungry wanderer came into a village. He filled an iron cauldron with water, built a fire under it, and dropped a 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 and others added potatoes, onions, carrots, mushrooms, and so on, until there was a meal for all.

1 The theses

One philosophical goal is *analysis*: the provision of necessary and sufficient conditions for a concept, or for the possession or application of a concept. The Western historical source of the goal is Plato's discussion of the concept of "virtue" in the *Meno*, but the *Meno* is also the source of a method: conjecture an analysis, seek intuitive counterexamples, reformulate the conjecture to cover the intuitive examples of the concept and to exclude the intuitive non-examples; repeat if necessary. Much of contemporary philosophy attempts the same strategy for many concepts: knowledge, belief, reference, causation, and so on. Addressing analyses of "reference," Mallon et al. (in press) argue that psychological investigation suggests that intuitions about reference are so varied that no uniform analysis can capture the discrepancies.

Our concern is about analyses of a scientifically and morally important notion, "actual causation"—about proposed necessary and sufficient conditions for one event to cause another. For an inference to a general analysis from intuitions about cases to be credible, more than psychological consensus is required. The intuitive cases used to justify an analysis must somehow be representative of the possible cases of actual causation or its absence. What is particularly interesting about "actual causation" is that the possible cases can in some sense be enumerated, and the enumeration can be used to show that consideration of intuitive examples is not representative, and apparently cannot be. Our argument first provides principles for enumerating the number of possible, structurally isomorphic examples of actual causal relations, without regard to the content of the related events. We show that even with very strong equivalence relations, and even considering only the number of events typical of examples in the philosophical literature, the number of possible cases is quite large. Second, we note that the number of equivalence classes grows exponentially as more events are considered. And, third, we show by example that as more events are added, novel kinds of ambiguous cases, or counterexamples to proposed analyses, emerge.

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, b), as well as of several other recent articles (Gilles 2005; Spohn 2005; Hiddleston 2005). Most of the literature is roughly Socratic and inductive: analyses are considered and a handful of "intuitive" story examples are considered in evidence. Some formal structure has been frequently imposed by reconstructing stories as Bayes net causal models: directed acyclic graphs (DAGs), with vertices that are variables and directed edges marking 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.[1] Within this framework, the various formal accounts of actual causation are justified by agreement of intuition; they are generally not derived from first principles, or justified pragmatically.

The graphical or "neuron diagram" representation permits a counting, or at least determination of lower bounds, of the number of inequivalent graphs, and thus, a lower bound on the number of possible different actual causation scenarios for any interpretation of the nodes of the diagram. Using the counts these representations permit, we argue that the inductive strategy for finding or testing a characterization of actual causation by intuitions about causal Bayes net cases may be futile because the number of cases potentially presenting distinct challenges to theories is unsurveyably large even with small numbers of potential causes. We consider a number of different restrictions on the space of possible cases, and argue that they are insufficient to make the problem tractable. One response might be to argue that cases with small numbers of variables suffice to determine the correct theory of actual causes. We argue, however, that interesting, novel distinctions and challenges arise for proposals for actual causation when we consider four- and five-cause systems, and so it is not plausible that all problems of interest are realized by cases with three or fewer potential causes. We further argue that the common graphical representation of actual causation is systematically ambiguous, largely because it ignores the potential relevance of the system state at previous times. We conclude with a positive proposal, using a different, less Platonic strategy: one can use causal Bayes nets (Spirtes et al. 1993) to unambiguously represent actual causal relations concerning changes from one initial state to another consequent on interventions, but this requires a significant shift in the typical understanding of actual causation. We do not claim the causal Bayes net framework solves all important questions about actual causation; most notably it does not address when absences are and are not causes. Finally, we consider the sparse but growing psychological literature on lay judgments of actual causation.

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] See, for example: Lewis (1986), Hitchcock (2001), Woodward (2003), and Halpern and Pearl (2000, 2005a), among many others.

(1) Events are represented by variables (usually taking two values but in principle without limit), possibly with one value (e.g., "0") marked for absences.
(2) Qualitative causal relations are represented by a directed acyclic graph (DAG) with the variables as vertices.
(3) Laws are given by deterministic or stochastic functions for each variable specifying its values as a function of the values (which may be probability distributions in the stochastic case) of its graphical parents. 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, the laws $A = B$, $C = f(B, A) = B \bullet A$, are jointly inconsistent with $A = 1$, $B = 0$, but $f(B, A)$ is defined for these values.[2]
(4) A realization of the system is an assignment of values to the variables, and a *legal* realization is an assignment consistent with the laws.
(5) A counterfactual realization α of realization ρ with respect to a proper subset $\mathbf{V}$ of the variables is a realization of the same system, differing from ρ for the variables in $\mathbf{V}$, assigning all variables that are not descendants of $\mathbf{V}$ their ρ values; the laws determine all other value assignments in α.

The scheme of definitions amounts to treating a causal counterfactual of the kind "if a had not happened ..." as an intervention, without backtracking, on a node whose values are *a* and *not-a*. An obvious variant would allow that counterfactual interventions in a stochastic system specify precise values for the variables directly intervened upon, rather than probability distributions.

In the deterministic case, with binary variables, these conditions amount to assuming an acyclic graphical causal model, in which the laws—the value of a child given its graphical parents—are given by truth functions. The same formalism lurks behind various probabilistic accounts of actual causation, only differing in making each child variable a stochastic function of the values of its parents. The intervention interpretation of counterfactuals (condition 5) is justified by two facts: (i) interventions satisfy the Lewis axioms for counterfactuals;[3] and (ii) 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).[4]

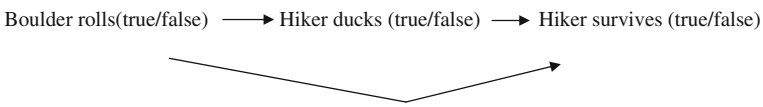
[2] Italicized upper case letters denote variables; lower case italized letters denote values of variables of the same lexicographic type; bold face words and letters denote sets. $\mathbf{A} \setminus \mathbf{B}$ denotes the members of $\mathbf{A}$ that are not members of $\mathbf{B}$.

[3] Lewis' axioms do not imply that A V B $\Box\!\!\rightarrow$ C |=A $\Box\!\!\rightarrow$ C V B $\Box\!\!\rightarrow$ C, which Pearl (2000) suggests is required by interventions (and Nute (1976) thinks is required by counterfactuals).

[4] For example, Lewis (1986), Hitchcock (2001), Woodward (2003), Hall (2004), Ramachandran (2004a, b), Kvart (2004a, b), Noordhof (2004), and Halpern and Pearl (2005a, b), Hiddleston (2005). 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,

Analyses of actual causation for deterministic cases have assumed that the relation obtains between *values* of variables representable in such networks. A method for determining actual causal relations thus depends on the actual values of variables, and truth functions for each directed edge in the graph. 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 shown below, and the central question can be expressed precisely as: "Is Boulder rolls = true, Hiker ducks = true, both, or neither, the actual cause of Hiker survives = true?"

Causal graph:



Laws:

Hiker ducks = Boulder rolls; Hiker survives = ~Boulder rolls v Hiker ducks.

Actual values:

Boulder rolls = true; Hiker ducks = true; Hiker survives = true.

Many cover stories obviously have the same formal structure; for example, "B throws a ball at a window S but H catches the ball" has the same structure as the Boulder/Hiker/Survival case. If we group together these obviously equivalent cases, then the philosophical literature discusses about a baker's dozen examples (see Sect. 4). Our first concern is whether this is an adequate sample of the number of possible cases, and whether an adequate sample of cases is possible at all (if each must be subjected to philosophical "intuition"). Consider the number of cases for the highly restricted case with: (i) 3 binary potential causes; (ii) 1 binary effect; and (iii) deterministic laws. Any subset of the three potential causes can be a cause of the effect (i.e., have a $C \rightarrow E$ edge in the graph), and so there are eight possible graphs: 1 graph with no causal connections; 3 graphs with one edge; 3 graphs with two edges; and 1 graph with all three edges. For each one-edge graph, there are 2^2 possible truth functions for the effect; for each two-edge graph, there are 2^4 truth functions; for the three-edge graph, there are 2^8 truth functions; and we treat the no-edge graph as just one case, since the truth "functions" are just constants. Altogether, there are 317 possible structures over the three potential causes and the effect.

This calculation does not yield all of the possible structures, however, since there can be causal relations among the potential causes (e.g., the Boulder $\rightarrow$ Hiker ducks connection). There are 25 distinct causal graphs over only the possible causes: 1 no-edge graph; 6 one-edge graphs; 6 three-edge graphs; 3 two-edge graphs in which both edges are directed into the same variable (a collider); and 9 other two-edge graphs. For each one-edge graph, there are 2^2 truth functions; for each two-edge graph—both collider graphs and non-colliders—there are 2^4 truth functions; and for each three-edge

Footnote 4 continued
seems to need an intervention account (which prevents backtracking) for some of his arguments (Hall 2004, pp. 261–262) but writes in terms of more general counterfactuals that allow them.

graph, there are $2^4 \times 2^2 = 2^6$ possible truth functions. If we again treat the no-edge graph as just one case, then there are 601 causal models over the three potential causes. Since any causal model among the potential causes can be paired with any structure for the effect, there are 190,517 possible causal models altogether. And the number of cases (not structures) is much larger: each possible structure corresponds to 2^C cases, where C is number of exogenous (i.e., no parent) variables in that structure. (Until further notice, we hereafter count only possible structures under various restrictions, bearing in mind that the number of cases will be much larger.[5]) Intuition obviously has too much to survey, and the standard cases clearly form an insufficient sample.

This analysis is of course a "worst-case" analysis: it assumed that the variable names matter (rather than just graphical structure), all possible laws/truth functions, and so on. One might hope that various natural restrictions on the set of possible structures could lead to a tractable number of cases; that hope will turn out to be in vain. The remaining parts of this section consider multiple plausible restrictions, and show that they are neither individually nor jointly sufficient to reduce the search space to a tractable size. We make no claims of completeness in this survey; there may be additional restrictions that would suffice, though we doubt that this is the case. This survey of plausible restrictions does, however, shift the burden of proof onto the proponent of a Socratic strategy, as the reliability of such a strategy depends directly on the number of possible causal models and cases.

2.1 Restricting the laws

Lewis (1986) restricted truth functions for his "neuron diagrams" to the form $E = (A_1 \mathrm{v} \cdots \mathrm{v} A_n) \,\&\, (\sim B_1 \& \cdots \& \sim B_k)$, but he did so only for the purposes of illustration, without any claim or suggestion that causal dependencies are so restricted. Cheng (1997) proposed that people use a psychological model of causation that, in the deterministic case, implies that the only causal models available to human judgement are isomorphic to neuron diagrams (Glymour 2003). Novick and Cheng (2004) subsequently considerably generalized this framework. Hiddleston (2005) adapts Cheng's (1997) earlier account to provide a theory of actual causation. For deterministic systems, his proposal yields the following: a causal model is a DAG with binary variables; for each variable Y, and each parent V_i of Y, the directed edge from V_i to Y is labeled either "generative" or else "preventative with respect to $V_j \cdots V_r$" where $\{V_j \cdots V_r\}$ is some other set of parents of Y. The value of Y is 1 if and only if at least one generative parent, V_k, of Y has the value 1 and no parent of Y that is preventative for Y for V_k has the value 1. A value $X = 1$ is an actual cause of a value $Y = 1$ if there is a directed path from X to Y such that every vertex on the path has value 1 and every edge is generative. Without notice or justification, Hiddleston's proposal excludes many elementary truth functions—exclusive *or* for example, and any truth functions that represent voting. For these and many other cases, edges cannot be unambiguously marked as "generative"

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

Table 1 Numbers of truth functions

Number of parents	Number of truth functions	Number of truth functions with test pairs
1	4	2
2	16	10
3	256	218
4	65,536	64,594
5	$>4 \times 10^9$	$>4 \times 10^9$

or "preventative." Understanding is not advanced by excluding, for no good reason, causal structures that are clearly possible and morally or scientifically relevant.

Restrictions on the truth functions should be based in general principles that are so central (but not necessary and sufficient) to the idea of causation that they need no inductive justification. The most natural such restriction is that a cause must (in some sense) actually matter for its effect in some condition. Every example of actual causation in the literature that uses graphical causal models to display the laws of the system implicitly uses a precise version of this restriction on truth functions:

(6) For each parent X of a variable Y, the function $Y = f(\mathbf{Parents}(Y))$ allows a *test pair* for X with respect to Y: two (not necessarily legal) realizations, α and β, such that (i) for all variables Z in $\mathbf{Parents}(Y) \backslash X$, $\alpha(Z) = \beta(Z)$; (ii) $\beta(X) \neq \alpha(X)$; and (iii) $f(\alpha(\mathbf{Parents}(Y)) \neq f(\beta(\mathbf{Parents}(Y))$.

The *test pair condition* is logically independent of the much discussed Markov property—the direct causes of a variable or event screen it off from variables or events that are not its effects—since the Markov condition formally allows a parent variable in a graph that is independent of its child. Given the Markov assumption, however, the test pair condition is implied by, but strictly weaker than, the Minimality condition (i.e., no proper subgraph of a graph satisfies the Markov condition for the probability distribution). Imposing the test pair condition reduces the number of allowable truth functions, but not much. Table 1 shows the counts.[6]

For a three-edge graph 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; there are thus 20 possible truth functions meeting the test pair condition for each three-edge graph. Similar reasoning for the different graphical possibilities yields 199 possible causal structures over three potential causes (see Table 2).

Consider now the ways that the effect variable can depend on the three potential causal variables in each of these 199 structures. There is 1 (trivial) test pair truth function for the no-edge case; 2 test pair truth functions for each of the three one-edge

[6] A closed form counting formula for truth functions satisfying the test pair condition for n variables is $\sum_{k=0}^{n} (-1)^k {}_nC_k 2^{(2^{n-k})}$. A counting formula, recursive in the number n of arguments, for truth functions meeting the test pair condition is: F(0) = 2, and $F(n) = 2^{2^n} - \sum_{i=1}^{n} \binom{n}{n-i} F(n-i)$.

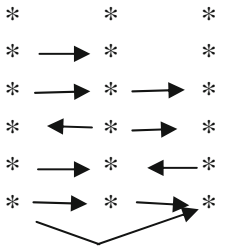
Table 2 Counting graphs with test pairs

Number of graphs of form…	× Number of test pair truth functions per graph	= Number of structures
1 disconnected graph	1	1
6 graphs of the form ->	2	12
6 graphs of the form ->->	$2 \times 2 = 4$	24
3 graphs of the form <- ->	$2 \times 2 = 4$	12
3 graphs of the form -> <-	10	30
6 graphs of the form -> ->	$2 \times 10 = 20$	120

cases (i.e., 6 possible structures); 10 permissible truth functions for the three two-edge cases (i.e., 30 possible structures); and 218 test pair truth functions for the single three-edge case. There are thus 255 possible structures (assuming the test pair condition) over the three potential causes and effect. 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 test pair restriction eliminates nearly 75% of the possible causal models, but that is not nearly reduction enough for intuition to survey the cases. Moreover, the combinatorics rapidly get much worse as the number of potential causes increases. The "simple" situation of five causes (i.e., all have $C \rightarrow E$) with *no* causal connections among them, and where we impose the test pair condition, corresponds to more than 4 billion possible structures.

2.2 Unlabeled graphs and other restrictions

We can additionally consider restrictions on the space of possible graphs. The idea with graphical models is that structure alone is considered, not the names given to variables or the substantive content of the events. In the absence of specific information about the meaning of variables, $X \rightarrow Y$ is structurally identical to $X \leftarrow Y$. If we group together directed acyclic graphs that are identical except for the variable names, then there are only six possible structures over three potential causes:



The middle column of Table 2 shows the number of test pair truth functions for each of these six graphs. The counts of structures involving the effect are more complicated if variable names do not matter. For example, if $X \rightarrow Y \leftarrow Z$ among the potential causes, then X, Y as the causes of E is equivalent to Z, Y being the causes of E; notice

Table 3 Counting unlabeled graphs with test pairs

	0 causes	1 cause	2 cause	3 cause	= Number of test pair truth functions for row structure
* * *	1	2	10	218	231
* → * *	1	3 × 2	3 × 10	218	255
* → * → *	1	3 × 2	3 × 10	218	255
* ← * → *	1	2 × 2	2 × 10	218	243
* → * → *	1	2 × 2	2 × 10	218	243
Three-edge	1	3 × 2	3 × 10	218	255

that X, Z being causes is not equivalent to the other two. Table 3 shows the number of test pair truth functions for E for all combinations of potential cause structure (rows) and number of causes of the effect (column). For cells with two numbers, the first number indicates the number of distinct graphical structures involving the effect when the potential causes are distinguishable only by their structural role (relative to the other potential causes).

We can compute the total number of causal models by multiplying the right-most column of Table 3 by the relevant number of test pair truth functions over the potential causes, and then summing together. Ignoring variable names, combined with the test pair condition, results in 10,263 possible causal structures. Smaller, but still a busy time for intuitions.

We can impose further plausible restrictions on the space of possible graphs, though they could conflict with some theories of actual causation.[7] All of the various accounts of actual causation agree that $C = \mathrm{c}$ 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 $B = b$. Various models are thus dispensable or equivalent with respect to testing an account of actual causation. For example, suppose E is a function of a single variable and * → * → * holds among the potential causal variables. The only distinct structure is the one in which E depends on the terminal star. If E depends on the middle variable, then it is equivalent to $* \rightarrow * \cdots *$ over the potential causes, since the last variable cannot be an actual cause, and cannot affect whether the other two variables are actual causes (by the above principles involving directed paths). If E depends on the first variable, then it replicates a case counted among those with $* \cdots * \cdots *$ as the relevant substructure on the causal variables. This restriction results in 20 distinct graphical structures over the three potential causes and E, distributed as shown in the central cells of Table 4. The relevant number of test pairs for the structures among the potential causes (rows) and involving the effect (columns) are also shown in Table 4.

[7] For example, that C is a cause of E if and only if C and $E = 1$ and the probability that $E = 1$ is higher given $C = 1$ than given $C = 0$.

Table 4 Counting unlabeled, minimal graphs with test pairs

TF multiplier for causes	Graph	Number of edges to E			Total graphs
		1	2	3	
1×	* * *	1	1	1	3
2×	*–>* *	1	2	1	4
4×	*–>*–>*	1	2	1	4
4×	*<–*–>*	0	1	1	2
10×	*–>*<–*	1	1	1	3
20×	*–>*<–*	1	2	1	4
TF Multiplier For edges to E		×2	×10	×218	

Table 5 Possible non-trivial truth functions

X	Y	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8	G_9	G_{10}
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

There are 9,682 causal models in all. Restrictions on the possible graphs and possible truth functions have significantly reduced the number of possible causal models, but there are still far too many to examine using intuition alone, and the number of models still grows super-exponentially with the number of variables.

2.3 Symmetries

Actual causation may also have symmetry relations that can be used to reduce the number of possible structures. Say that one truth function G_i is a *value negation* of another G_j if they have the same argument variables and, for all valuations of the argument variables, G_i and G_j have opposite outputs. For example, Table 5 shows the 10 test pair truth functions for two arguments; the value negation partition classes are: $\{G_1, G_2\}$, $\{G_3, G_4\}$, $\{G_5, G_6\}$, $\{G_7, G_8\}$, and $\{G_9, G_{10}\}$. For every value of X, Y, G_1 and G_2 have opposite values, and so each is a value negation of the other.

Value negation obviously defines a partition of the set of truth functions into two-member classes, and it preserves the test pair condition. If we consider truth functions to be equivalent to their value negations, then the number of test pair truth functions is cut in half: there are 5 functions for 2 arguments, 109 for 3 arguments, but, unfortunately, more than 2 billion for 5 arguments.

We might further assume that actual causation is symmetric with respect to interchange of true and false in all arguments: truth functions G_i and G_j are *argument negation equivalent* if $G_i(X) = G_j\left(X_{\mathrm{t/f}}\right)$ for all valuations of X, where $X_{\mathrm{t/f}}$

Table 6 Example of truth function invariance under variable permutation

X	Y	$Z = \mathrm{G}_1$	$Z = \mathrm{G}_2$
T	T	F	F
F	T	T	F
T	F	F	T
F	F	T	T

substitutes F for T and T for F in X. Equivalence under argument negation also preserves the test pair property and partitions the truth functions. The resulting classes for functions of two variables meeting the test pair condition (Table 5) are: $\{G_1, G_7\}, \{G_2, G_8\}, \{G_3, G_5\}, \{G_4, G_6\}, \{G_9\}$, and $\{G_{10}\}$. This 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 as equivalent any truth functions that are equivalent under value negation or argument negation. This yields a three-class partition of the test pair truth functions: $\{G_1, G_2, G_7, G_8\}$, $\{G_3, G_4, G_5, G_6\}$, and $\{G_9, G_{10}\}$. Combined with the other restrictions, the number of cases for three potential causes begins to seem surveyable. These symmetry restrictions do not, however, change the fundamentally exponential growth of the number of acceptable truth functions.

One further potential formal symmetry principle deserves remark. Consider the structure $X \rightarrow Z \leftarrow Y$, and the two possible truth functions shown in Table 6. Since the labels of variables are not meaningful, we might argue that G_1 and G_2 in Table 6 are really the same truth function, since G_1 becomes G_2 when we permute the X, Y values; that is, $G_1(Y, X) = G_2(X, Y)$ and $G_1(X, Y) = G_2(Y, X)$. Any permutation of argument columns in the truth tables takes a 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 $N = 2$, for example, permuting arguments results in no additional reduction of classes. Nonetheless, unlike the other restrictions, permutation equivalence yields an exponential (as a function of N) reduction in the number of "equivalent" truth functions. But, again, the number of test pair cases grows super-exponentially. Thus, for 5 arguments, 4 billion plus truth functions are reduced to about 35 million 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 is 1 class for $N = 1$; 3 classes for $N = 2$; 26 classes for $N = 3$; and 1,579 classes of allowable truth functions for $N = 4$. These counts are only for the number of truth functions; we must again consider all of the different graphical structures, and then determine the number of cases for each possible structure. The number continues to grow exponentially in the number of variables. All of these restrictions have helped, but they are not sufficient to make a extensional Socratic strategy viable.

As we suggested above, the proponent of a Socratic strategy might propose still more restrictions, though we doubt that this strategy will ultimately prove worthwhile. A different way to save the Socratic strategy would be to argue that cases with few variables suffice. That is, extra variables present nothing new, and so the exponential growth is irrelevant. We disprove that line of response by examples, but doing so requires consideration of particular accounts of actual causation. We thus detour in the next section to consider two recent theories.

3 Four theories and their examples

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 additional 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 Glymour and Wimberly (2007) and in Glymour (2005), but our point is simply that interesting cases arise for four and more potential causes. The combinatoric explosion cannot be avoided.

Building from earlier proposals by several authors, James Woodward (2003, pp. 83–84), makes the following proposal.

W: "Consider a particular directed path P from X to Y and those variables $V_1 \ldots V_n$ that are not on P. Consider next a set of values $v_1 \ldots v_n$, one for each of the variables V_i. The values $v_1 \ldots v_n$ 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 $v_1 \ldots v_n$ that will change the actual value of Y. . . .

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.[8]

"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 Z_i 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 Z_i.

[8] Eric Hiddleston (in commentary on the paper at FEW 2007) suggested that the lexicographic ordering over AC and AC′ should be applied on a per-variable basis (i.e., for each variable, if AC does not apply, then use AC′), rather than over all variables as we do here (i.e., if AC applies for *any* variable, then stop). The relevant passages in Woodward (2003) are ambiguous, but Woodward (personal communication) indicated that our interpretation was his intended account.

Halpern and Pearl (2005a) have recently made a different proposal:

> **HP2005**: "(M, **u**) |= $[X \leftarrow x]\phi$" abbreviates 'ϕ is true in structure M for legal realization **u** if **u** is possibly altered by an intervention setting X to value x.'
> $\mathbf{X} = \mathbf{x}$ is an actual cause of ϕ in (M, **u**) if and only if:

AC1: (M, **u**)|= ($\mathbf{X} = \mathbf{x}$) and ϕ

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

(a) (M, **u**)|= $[\mathbf{X} \leftarrow \mathbf{x}', \mathbf{W} \leftarrow \mathbf{w}']\sim\phi$

(b) (M, **u**)|= $[\mathbf{X} \leftarrow \mathbf{x}.\mathbf{W}' \leftarrow \mathbf{w}', \mathbf{Z}' \leftarrow \mathbf{z}^{\bullet}]\phi$

for all subsets $\mathbf{W}'$ of **W** and for all subsets $\mathbf{Z}'$ of **Z**. In words, setting any subset of variables in **W** to their values in $\mathbf{w}'$should have no effect on ϕ, as long as **X** is kept at its current value **x**, even if all the variables in an arbitrary subset of **Z** are set to their original values in the context **u**.[9]

AC3: **X** is minimal; no [proper] subset of **X** satisfies conditions AC1 and AC2.

AC4: $\mathbf{X} = \mathbf{x}$ and $\sim \phi$ is consistent."

Both of these accounts of actual causation are, in part, justified by their fit with our intuitions on salient cases. Failures (by simpler versions) to fit our intuitions are responsible for much of the complexity in both of these accounts. This Socratic strategy is largely driven by a relatively standard set of stories. Each of these cases—14 presented below, but variants are all over the literature—corresponds to a set of truth functional relations among propositional variables and a valuation of the variables. The cover stories are (or should be) irrelevant, as the truth functional relations and valuations can be realized with switches and lights in electrical circuits, or indeed in any present-day computer. We provide the W and HP2005 predictions for each example.

(1) A and B each fire a bullet at a target, simultaneously striking the bullseye (D). What caused the bullseye to be defaced?

$$A \rightarrow D \leftarrow B \quad D = A + B; A = B = D = 1$$

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

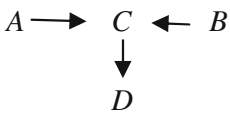
(2) A and B each fire a bullet at a target. A's bullet travels faster, knocking out the bullseye (D), which B's bullet would have knocked out a moment later (D') otherwise. What caused the event $D = 1$, of the bullseye's removal?

$$B \rightarrow D' \leftarrow A \rightarrow D; D = A, D' = B(1 - A); A = B = D = 1; D' = 0$$

W, HP2005: The actual cause of $D = 1$ is $A = 1$.

[9] 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.

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



$C = A \bullet B; \quad D = C; \quad A = B = C = D = 1.$

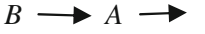
W, HP2005: The actual causes of $D = 1$ are $A = 1$, $B = 1$, and $C = 1$.

(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?

$$C \rightarrow B \rightarrow A \rightarrow D; \quad D = A; \quad A = (1 - B); \quad B = (1 - C);$$
$$A = D = C = 1, B = 0.$$

W, HP2005 : The actual causes of $D = 1$ are $A = 1$, $C = 1$, and $B = 0$

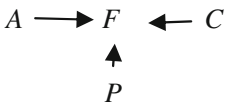
(5) A, an imperfect marksman, is about to fire at the target, but his aim is too low. B standing at the back of the crowd, could push his way through to A and lift the rifle barrel just the right amount, but B does no such thing. A's bullet misses the bullseye. What caused the bullseye to be missed ($D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



$D = (1 - A), A = (1 - B). \quad B = 0, A = 1, D = 0.$

W, HP2005: The actual causes of $D = 0$ are $B = 0$ and $A = 1$

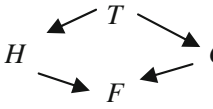
(6) A, the perfect marksman, aims (A) at the target, but fails to cock his gun (C), and pulls the trigger (P). The gun does not fire and the target is untouched. Which event caused the gun not to fire ($F = 0$)?



$D = A \bullet C \bullet P = F; \quad A = P = 1; \quad C = F = 0.$

W, HP2005: The actual cause of $D = 0$ is $C = 0$

(7) A 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 ($F = 0$)?

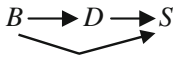


$H = (1-T), C = T, F = H \bullet C; T = C = 1, H = F = 0.$

W, HP2005: 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?
$B \rightarrow H \rightarrow D$
$H = 2$ 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, HP2005: $B = 1$ caused the left-handed shot ($H = 2$)
W, HP2005: $H = 2$ caused the bullseye hit ($D = 1$); $B = 1$ did not cause it

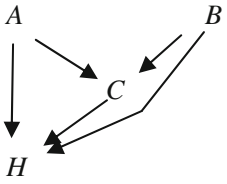
(9) A boulder slides ($B = 1$) toward a hiker, who, seeing it, ducks ($D = 1$). The boulder misses him and he survives ($S = 1$). Did the boulder sliding cause his survival?



$S = (1 - B) + D; \quad S = B = 1$

W, HP2005: 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).



$C = B \bullet A, H = B \bullet (1-C)+A. \quad A = B = H = C = 1$

W, HP2005: the actual cause of $H = 1$ is $A = 1$

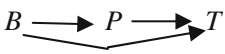
(11) A and B, both perfect marksmen, pull their triggers on similar guns at the same time. B loaded her rifle ($Lb = 1$) and hits the bullseye ($H = 1$). A has forgotten to load his rifle ($La = 0$). What caused the hit?

$$H = A \bullet La + B \bullet Lb \quad A = B = Lb = H = 1; La = 0$$

W, HP 2005: The actual causes of $H = 1$ 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 ($B = 1$) which prevent a pregnancy ($P = 0$) that, had it occurred, would have caused the actual thrombosis ($T = 1$) caused by taking the birth control pills.



$P = (1 - B)$; $T = B + P.B = T = 1$, $P = 0$

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. A and B both vote for $C(A, B = c)$.
W: $A = c$ is the actual cause of C winning .
HP2005: $A = c$, $B = c$ are the actual causes of C winning.

W and HP agree in every case except 14, where W's judgement seems (to us) the more plausible. In this case, and in others such as case 10, the analysis of the W account depends critically on stopping when condition AC is satisfied. There are alternative, simpler proposals that agree with the judgments of both W and HP2005 on many of these cases, but differ on others. As examples, consider the following two proposals, neither of which we endorse.

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 $y' \neq y$ for Y, and X is a member of a set $\mathbf{X}$ of variables (not having Y as a member, of course) with actual values $\mathbf{x}$, and there exist alternative values $\mathbf{x}'$, none of which equal the corresponding values in $\mathbf{x}$, such that an intervention on the system in state s that fixes $\mathbf{X} = \mathbf{x}'$ entails $Y = y'$, and no proper subset $\mathbf{Z}$ of $\mathbf{X}$ with actual values $\mathbf{z}$ is such that there exist alternative values $\mathbf{z}'$, none of which equal the corresponding values in $\mathbf{z}$, such that an intervention on the system in state s that fixes $\mathbf{Z} = \mathbf{z}'$ 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 $\{A = 1, B = 1\}$: if A and B are changed 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.

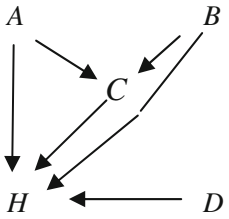
4 Looking further

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. Cases 8 and 14 above, the latter of which separates the two main proposals, already involve variables with three values. We will show that new actual causation "phenomena" occur when more potential causes are allowed, in cases with three valued variables and in cases with only binary variables. We cannot say how many such novel puzzles there are, and that is the point.

Consider Woodward's account W of actual causation. For any system in which AC1 and AC2 of W apply, we can make them *not* apply (and so force consideration

of AC′1 and AC′2) by the introduction of a new, overdetermining cause, and thereby change judgments about variables. In particular, an event that is not an actual cause in some scenario can become one simply through the addition of an additional, completely unrelated cause. As an example, consider case 10, but now with an extra, overdetermining cause D.

(10a) A and B, and D, all 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 at the same time as D's bullet. What caused the bullseye to be hit ($H = 1$)?



$$C = B \bullet A,\ H = B \bullet (1 - C) + A + D$$
$$A = B = D = H = C = 1$$

In the original case 10, $A = 1$ was the only actual cause of $H = 1$ on all of the analyses; $B = 1$ was not an actual cause, which is intuitively correct since A's shot preempts B. When D is added to the system, however, AC1 and AC2 of W no longer apply, and so we must apply AC′1 and AC′2. The redundancy set $\{D = 0, C = 0, A = 0\}$ for the $B \rightarrow H$ path then implies that $B = 1$ is an actual cause of $H = 1$ in case 10a. Case 14 behaves similarly if an extra potential cause, F, is added with two possible values ($F = c$, or no vote), and where $F = c$ is sufficient for C to win. If actually $F = c$, then $B = c$ becomes an actual cause of C's victory. It is easy to see why adding an additional cause might change an actual cause into a non-cause; this instability is troubling precisely because it involves a non-cause becoming an actual cause.[10] This instability cannot be seen, however, in structures with only three variables.

Cases with more variables create new difficulties for HP2005 as well. Consider an example with five variables.

(15) A ranch has five individuals: Cowboy C, Ranger R, Wrangler W, and two Hands H_1, H_2. Everyone votes either for staying around the campfire (0), or for going on a round-up (1). A complicated rule is used to decide the outcome O: (a) if $C = R$, then $O = R$; (b) if R differs from the other four, then $O = R$; and (c) otherwise, majority rules. Suppose $C = R = 1$ and $W = H_1 = H_2 = 0$ (and so $O = 1$). Was $W = 0$ an actual cause of $O = 1$?

We need first to consider whether W's vote might be 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 W wanted not to go on a round-up and W is in ignorance about how all of the

[10] This kind of problem seems apposite since various forms of stability were among the criteria Woodward considered for causal relations.

others will vote: his priors for every vote but his are 50/50 for round-up. No matter how W votes, cases in which C and R agree on 0 (= stay by the campfire) are equally likely as cases in which C and R agree on 1 (= go on a round-up). Averaged over these cases, W 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 the other four individuals. In two of these patterns, R stands alone and W has an equal chance (averaged over these cases) of getting his desire if he votes 0 as if he votes 1. Ignore them. There remain six cases in which R and C do not agree and R does not stand alone (ignoring W'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 W votes 0, then $O = 0$ in the first five cases; if W votes 1, then $O = 0$ in the 3rd row only. Thus, $W = 0$ is a vote against a round-up; the H_1 and H_2 votes are similarly non-strategic. Nonetheless, the actual causes according to the various proposals are:

W, SimpleJ: $R = 1$ is the only actual cause
HP2005, Simple: $R = 1$; $W = 0$; $H_1 = 0$; $H_2 = 0$ are all actual causes

Things come apart in a novel way in this case.[11] What perplexities lurk elsewhere among the manifold unexamined examples?

5 Whose judgment?

Even if we somehow solved the combinatoric explosion, there is reason to be concerned about the reliability of the Socratic strategy. The success of that strategy (if any) will greatly depend on the relative stability of the relevant intuitive judgments. All instances of the Socratic strategy that we know rely on judgments of a small group of philosophers, even for unusual cases. The presumption that philosophers' judgments 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 about a range of cases—some simple, some more complex—and how they differ (if at all) from philosophers' judgments. More radically, one would like to know what proportion of informed individuals would reject as ambiguous the very question of actual causation in one or another

[11] 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.

description of circumstances. One would like to know whether judgments of actual causation depend only on the final state or on the transitions that lead to it. One would like to know in what respects systems are sometimes too complex for people to give more than random judgments, or none at all. And many other questions remain unanswered.

There is an enormous psychological literature on human judgment 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), but relatively little about actual causation in other contexts. 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 judgment imbedded in morally fraught contexts (e.g., Ahn and Kalish 2000; Ahn et al. 1995; Wolff and Song 2003), and in social contexts. In particular, Choi et al. (1999) focused on causal attributions in a variety of social situations by participants in a range of cultures. Their major conclusion was that participants in Asian cultures are more inclined towards situationism: they are more likely to attribute people's (token) actions to situations, rather than dispositional or personality traits of the individual.

There is an even more limited psychological literature on the kinds of cases philosophers have considered. Perhaps the most relevant piece of psychological work is Walsh and Sloman (2005). They provided experimental participants with a range of "standard cases" from the philosophy literature, including overdetermination, late pre-emption, and interruptions (A is going to cause E but B intervenes by blocking A; did A cause E not to happen? Did B cause E not to happen?). Their results were decidedly ambiguous: except in the clearest cases—those on which the entire philosophical community agrees—the modal description for each 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 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. No similar study of philosophical cultures is yet available.

6 Misrepresentation and metaphysics

There is little reason to expect a Socratic strategy to succeed in finding a correct theory of actual causation; there are too many cases, and intuitive judgments about the cases are almost certainly too unstable. Rather than trying to find necessary and sufficient conditions for actual causation, a "Euclidean" strategy aims to provide reliable

indicators for discovering actual causal relations. Those indicators might provide a definition of actual causation, but they need not. The justification of a Euclidean account of actual causation is provided by its fruitfulness in generalization, inference, control, and so forth. Perfect fit with intuition, or applicability in all possible situations, are not desiderata for a Euclidean account, precisely because it does not try to provide necessary and sufficient conditions. There are of course more and less sophisticated versions of the Euclidean strategy, depending on both the scope of application, and the criteria for assessing fruitfulness.

We suggest progress can be made by adopting a Euclidean strategy that searches for reliable indicators of actual causation. In many situations, *changes* in variable values are identified as actual causes, or at least the important actual causes. Changes in variable value might be neither necessary nor sufficient for something being an actual cause, but that is the wrong standard for a Euclidean strategy.

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. It includes a characterization, given a causal model, of the effects of exogenous *changes* in any set of variables on any disjoint set of variables. That characterization, titled the Manipulation Theorem, made no reference to a *static* set of variable values causing some other set of values. The various adaptations of Bayes nets for descriptions of actual causal relations all attempt to introduce exactly such a relation between static states; alternative states are only referenced in the counterfactual or intervention conditions of the analyses. Halpern and Pearl (2005a, n. 6) explicitly state that 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." And that is part of the problem. None of the graphical model accounts of actual causation include changes over time. 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 some of the philosophical literature on actual causation.

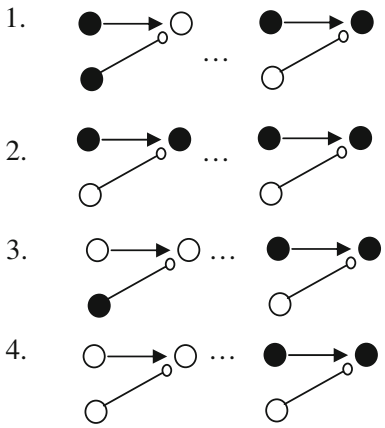
We tend to think of causes as changes, or happenings, but the reading of causal graphs that we have considered so far does not encode this information. A distinguished value (say "1") represents simply the occurrence of some event, and the other value (e.g., "0") represents the *absence* of that event. The event itself might be a change of some feature of a space time-region, but it can equally well be the continuation of an enduring condition. 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.[12] 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

[12] *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.

previous events. Even in the absence of such problems, this reading of causal graphs fails to capture the importance of changes.

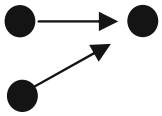
For Lewis-style "neuron diagrams," actual causal relations typically involve at least two total states over time, each specifying values for the variables of the system. The between-state changes of values for some variables bring about a change in other variables, or prevent changes in other variables that would otherwise have occurred. In a single diagram with a single value for each vertex (i.e., a single time-slice), intuitions about what causes what may vary because people implicitly make different assumptions about the prior states.[13] Outside of formal representations, this prior state information gets glossed as "normal conditions" or "the causal field" or perhaps "defaults." In discussions of actual causation, it is generally left inexplicit, but informality is not a solution to equivocation.

Consider a system of three nodes/variables that changes over time. Now consider four possible transitions (from left to right) in the system state, 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.



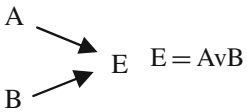
In all four cases, the final state is the same, but we wager that many people, shown the sequences—or their equivalents in some less abstract representation of the same structures and state relations—would not judge the causes of the final state of the right-most node to be the same in all four cases. We expect that common judgments would locate the cause in sequences #1, 3, and 4 to be the state changes in 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 is an interesting and open question. This focus on changes suggests representing the value changes themselves as nodes, which provides a different representation of each of the above sequences. For example, the *changes* in the third sequence above might be represented as:

[13] 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 and Nelson 2003; Sher and McKenzie 2006).

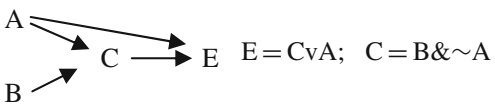


where a dark node indicates that a local change of state occurred. But the change (or happening) graph representation has no clear functional dependencies that are independent of the actual beginning and end states—no laws—and fails to mark the difference between a change in a node from empty to dark, and a change of that same node from dark to empty; each kind of change becomes a dark node. The same "change graph" would also represent this transition:

There are other ambiguities that can arise 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 can become A→ pretty-much-anything-you-want→B. The probability relations, intervention relations, and variable causation between A and B all remain unaltered. But in some cases the actual causation relations are arguably changed. Hall (2004) has pointed out that the diagram and truth function:



is consistent with the mechanism:



When A = B = E = 1, B is arguably an actual cause of E = 1 in the first causal model, but is less obviously an actual cause in the second causal model.

These problems vanish if we consider changes produced by exogenous changes in a particular system state. 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 is assumed in all of the discussions we have mentioned. Strengthening the Markov assumption with Minimality—which implies the test pair condition— then permits an algorithm for computing the changes an ideal intervention produces (Pearl 2000). On this Euclidean approach based in state changes, no induction over cases is required and various problems (e.g., Hall's) disappear: given the state of the system, and an intervention on a variable (or variables), the resulting changes in other variables' states are unambiguous.

Puzzles of course remain beyond those ambiguities inevitable in a formal representation of informal language. For example, we need an account of causal explanations of non-changes by combinations of changes and non-changes, such as "the rains did not flood the valley because the dam did not break."

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 generally to be isomorphic to 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. We are not arguing for abandoning neuron diagrams or Bayes nets or graphical causal models in philosophical investigations of causal relations. We are not arguing against the possibility of a correct theory of actual causation. It is instead an argument (i) against the adequacy of the unsystematic Socratic strategy that has dominated philosophical discussion of actual causation; (ii) against the sufficiency of Bayes net representations for actual causation without consideration of state transitions; and (iii) 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. doi:10.1016/0010-0277(94)00640-7.

Cheng, P. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104*, 367–405.

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

Collins, J., Hall, N., & Paul, L. (Eds.). (2004). *Causation and counterfactuals.* Cambridge, MA: MIT Press.

Dowe, P., & Noordhof, P. (Eds.). (2004). *Cause and chance: Causation in an indeterministic world*. New York: Routledge.

Gilles, D. (2005). An action-related theory of causality. *The British Journal for the Philosophy of Science, 56*, 823–842. doi:10.1093/bjps/axi141.

Glymour, C. (2003). *The mind's arrows: Bayes nets and graphical causal models in psychology*. Cambridge, MA: MIT Press.

Glymour, C. (2005). Review of "Causality and chance" edited by D. Dowe & P. Noordhoff. *Mind*, *114*(455), 728–733.

Glymour, C., & Wimberly, F. (2007). Actual causation and thought experiments. In J. K. Cambpell, M. O'Rourke & H. Silverstein (Eds.), *Causation and explanation* (pp. 43–68). Cambridge, MA: MIT Press.

Hall, N. (2004). Two concepts of causation. In J. Collins, N. Hall & L. Paul (Eds.), *Causation and counterfactuals*. Cambridge, MA: MIT Press.

Halpern, J., & Pearl, J. (2000). *Causes and explanations: A structural-model approach.* Technical report R-266. Cognitive Systems Laboratory. University of California at Los Angeles.

Halpern, J., & Pearl, J. (2005a). Causes and explanations: A structural-model approach. Part I: Causes. *The British Journal for the Philosophy of Science, 56*, 853–887.

Halpern, J., & Pearl, J. (2005b). Causes and explanations: A structural-model approach. Part II: Explanations. *The British Journal for the Philosophy of Science, 56,* 889–911.

Hiddleston, E. (2005). Causal powers. *The British Journal for the Philosophy of Science, 56*, 27–59. doi:10.1093/phisci/axi102.
Hitchcock, C. (2001). The intransitivity of causation revealed in equations and graphs. *The Journal of Philosophy, 98*, 273–299. doi:10.2307/2678432.
Kvart, I. (2004a). Probabilistic cause, edge conditions, late preemption and discrete cases. In P. Dowe & P. Noordhof (Eds.), *Cause and chance: Causation in an indeterministic world*. New York: Routledge.
Kvart, I. (2004b). Causation: Probabilistic and counterfactual analyses. In J. Collins, N. Hall & L. Paul (Eds.), *Causation and counterfactuals*. Cambridge, MA: MIT Press.
Lewis, D. (1986). Causation. In *Philosophical Papers*, Vol. II, New York: Oxford University Press.
Mallon, R., Machery, E., Nichols, S., & Stitch, S. (in press). Against arguments from reference. *Philosophy and Phenomenological Research*.
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 & Review, 10*, 596–602.
Menzies, P. (2004). Difference making in context. In J. Collins, N. Hall & L. Paul (Eds.), *Causation and counterfactuals*. Cambridge, MA: MIT Press.
Michotte, A. (1954). *La perception de la causalité*. Louvain: Publications Universitaires de Louvain.
Noordhof, P. (2004). Prospects for a counterfactual theory of causation. In P. Dowe & P. Noordhof (Eds.), *Cause and chance: Causation in an indeterministic world*. New York: Routledge.
Novick, L. R., & Cheng, P. W. (2004). Assessing interactive causal influence. *Psychological Review, 111*, 455–485. doi:10.1037/0033-295X.111.2.455.
Nute, D. (1976). David Lewis and the analysis of counterfactuals. *Nous (Detroit, Mich.), 10*, 455–461. doi:10.2307/2214616.
Pearl, J. (2000). *Causality*. New York: Oxford University Press.
Ramachandran, M. (2004a). Indeterministic causation and varieties of chance raising. In P. Dowe & P. Noordhof (Eds.), *Cause and chance: Causation in an indeterministic world*. New York: Routledge.
Ramachandran, M. (2004b). A counterfactual analysis of indeterministic causation. In J. Collins, N. Hall & L. Paul (Eds.), *Causation and counterfactuals*. Cambridge, MA: MIT Press.
Sher, S., & McKenzie, C. R. M. (2006). Information leakage from logically equivalent frames. *Cognition, 101*, 467–494. doi:10.1016/j.cognition.2005.11.001.
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., Glymour, C., & Scheines, R. (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. doi:10.1093/bjps/axi151.
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. doi:10.1016/S0010-0285(03)00036-7.
Woodward, J. (2003). *Making things happen: A theory of causal explanation.* Oxford: Oxford University Press.