6

Causal Asymmetry from Statistics

Seth Lloyd†
California Institute of Technology
Pasadena, CA 91125, USA

6.1 Introduction

Aristotle's analysis of the world in terms of cause and effect formed a cornerstone of Renaissance thought. Efficient cause explained the sinfulness of the world as an effect of Adam's fall, while final cause justified religious and political institutions as 'caused' by God's intentions for the future of mankind. The Enlightenment's questioning of religious and social institutions robbed arguments by final cause of much of their force. Hume denied final cause, and regarded efficient cause as arising simply from the human habit of calling 'cause' the first in time of two events that occur in constant conjunction (Hume 1739). By the beginning of the current century, the intellectual status of causal reasoning had receded to the point that Russell could write (Russell 1929), "The law of causality, I believe, like much that passes muster among philosophers, is a relic of a bygone age, surviving, like the monarchy, only because it is erroneously supposed to do no harm."

Russell's conviction of the anachronistic nature of causal law came from well–established 19th century ideas in physics: the fundamental description of the world was given by Hamiltonian evolution; the behaviour of Hamiltonian systems over a given time interval was determined equally by conditions given at the beginning or at the end of the interval; the particular Hamiltonians that seemed to describe physical systems were quadratic in momentum, and so were invariant under the transformation $t \longrightarrow -t$. In this view, the underlying dynamics of the world were completely time symmetric, and temporal and causal asymmetry were merely artifacts of our inability to perceive the true microscopic workings of the world around us, an inability that forced us to rely on statistical descriptions.

The greater the explanatory power of a prevalent physical model, the greater is the temptation to regard that model's picture of the world as fundamental. Russell's dismissal represents a low water mark in causal and statistical reasoning. In the twentieth century, the successes of statistical and quantum mechanics have

† This work supported in part by the U. S. Department of Energy under Contract No. DE-AC0381-ER40050

lent legitimacy to probabilistic reasoning, to the extent that most physicists now regard the deterministic evolution of classical mechanics as an approximation, and the stochastic natures of quantum mechanics and of chaos as fundamental. With the growing importance of statistical techniques has come a resurgent interest in causality and its relationship to statistics, exemplified by the work of Reichenbach (Reichenbach 1956). Causal reasoning, like the monarchy, survived both Russell and his remark.

In the present work, the only requirement made of a cause is that under some circumstances, variation in the outcome of the cause produces a correlated variation in the outcome of the effect. In essence, the cause–effect relationship can be thought of as a one–way communications channel through which the cause sends information that the effect receives. The basic idea here is that causal connection implies the possibility of statistical correlation, while the absence of causal connection implies the absence of correlation. If two events are correlated, then either one has an effect on the other, or the two have a common cause in their past. By extending these ideas to many events, one arrives at a method for deriving patterns of statistical dependence and independence from the causal relations between the events.

6.2 Correlation and Information

This programme may be made formal as follows. Suppose that the statistical information possessed about a set of events $A, B, C, \ldots$ takes the form of a joint probability distribution $p(abc\ldots)$ over the possible outcomes $a, b, c, \ldots$ of $A, B, C, \ldots$. From this joint distribution, one can derive various marginal distributions, such as $p(ab) \equiv \sum_{c\ldots} p(abc\ldots)$, the probability for A to have outcome a and B to have outcome b, and conditional probability distributions, such as $p(a|b) \equiv p(ab)/p(b)$, the probability that the outcome of A is a given that the outcome of B is b.

Two events A and B are correlated if fixing the outcome of B can change the probabilities for the outcome of A, that is, if $p(a|b) \neq p(a)$ for some a, b. Note that $p(a|b) \neq p(a)$ if and only if $p(b|a) \neq p(b)$: correlation is a symmetric relationship. On the one hand, if you throw a rock at a window, there is a better than normal chance that the window will break. On the other hand, if a window breaks, there is a better than normal chance that someone has thrown a rock at it. The degree of correlation between events can be measured using information theory (Shannon and Weaver 1949). Define $I(A) \equiv -\sum_a p(a) \log_2 p(a)$; $I(A)$ is the average number of bits of information required to specify the outcome of A. If the outcome of B is fixed, the average number of bits required to specify the outcome of A is $I(A|B) \equiv \sum_b p(b)\big(-\sum_a p(a|b) \log_2 p(a|b)\big) = I(AB) - I(B)$, where $I(AB) = -\sum_{ab} p(ab) \log_2 p(ab)$ is the average number of bits required to specify the outcome of both A and B. The degree of correlation between A and B can be measured by how much knowledge of the outcome of B reduces the amount of information required to specify the outcome of A: the corresponding quantity,

$I(A;B) \equiv I(A) - I(A|B) = I(A) + I(B) - I(AB)$, is called the mutual information between A and B. Note that $I(A;B) = I(B;A)$: the amount that one finds out about A by knowing B is equal to the amount that one finds out about B by knowing A. Note also that $I(A;B) \geq 0$, with equality if and only if $p(a|b) = p(a)$ for all a, b. If $I(A;B) = 0$, then A and B are said to be independent: knowing the outcome of B imparts no knowledge about the outcome of A.

Mutual information was originally defined in communications theory, to measure the capacity of communication channels. Regarding the cause–effect relationship as such a channel, one may take the mutual information between cause and effect as a measure of the amount of information that the effect is receiving from the cause. In fact, this is strictly true only when an event has only one cause. When an effect has more than one cause, just as when a communications channel has more than one input, one must be more careful in measuring the amount of information transmitted from cause to effect.

6.3 Causal Models

Directed graphs will be used to model the causal relationships between events. So, for example, $A \longrightarrow B$ will indicate that A has an effect on B in the sense given above, that variation in the outcome of A induces a correlated variation in the outcome of B under some circumstances. Similarly, $A \longrightarrow B \longrightarrow C$ indicates that A has an effect on B, B has an effect on C, and that A may have an effect on C, but only through its effect on B. That is, $A \longrightarrow B$ indicates that A has a *direct* effect on B, unmediated by any of the other variables about which we possess statistical information.

6.3.1 Models with Two Variables

Causal models imply the existence of independence relationships between the events in the model. Consider a model, $A \quad B$, for the events A and B: in this model, there is no causal connection whatsoever between A and B. Since there is no casual connection between A and B, there should be no correlation between A and B: that is, the model, $A \quad B$ implies that $I(A;B) = 0$. Similarly, the models, $A \longrightarrow B$ and $B \longrightarrow A$, both imply that $I(A;B)$ need not be equal to zero.

At first sight, it might be thought that $A \longrightarrow B$ should imply that $I(A;B)$ is strictly greater than zero. This is not the case. Consider, for instance, an exclusive OR gate, whose output is equal to 1 when exactly one of its inputs is 1, and equal to 0 otherwise. Label the inputs A, C and the output B. Suppose that each of the four possible combinations for the inputs A and C, 00, 01, 10, and 11, have equal probability. It is easy to verify that even though variation in the input A can cause a correlated variation in the output B, in the absence of knowledge of the value of C, A and B are uncorrelated. The reason for this lack of correlation is that the

value of C determines the way in which variation in A induces variation in B: if $C = 0$, then variation in A produces a perfectly correlated variation in B — when A is 0, B is 0, when A is 1, B is 1; if $C = 1$, then variation in A produces a perfectly anticorrelated variation in B — when A is 0, B is 1, when A is 1, B is 0. If both values for C are equally probable, then the correlation between A and B when $C = 0$ is counterbalanced by the anticorrelation when $C = 1$, and A is not correlated with B, even though it is a cause of B. If $C = 0$ and $C = 1$ are not equally probable, however, A and B will be correlated. $A \longrightarrow B$ generally implies that $I(A;B) > 0$; special conditions are required on the other causes of B to destroy this correlation.

For causal models containing only two variables, the situation can be summarized as follows: $A \quad B$ implies that $I(A;B) = 0$; $A \longrightarrow B$ and $B \longrightarrow A$ generally imply that $I(A;B) > 0$, but are also consistent with $I(A;B) = 0$. If the statistics for the outcomes of A and B give $I(A;B) > 0$, then they rule out the model $A \quad B$. $I(A;B) > 0$ does not imply that either $A \longrightarrow B$ or $B \longrightarrow A$ is the case, however. Correlation between A and B can also be explained by the existence of a common cause $\mathscr{X}$ that lies outside the set of events about which we possess statistics: $A \longleftarrow \mathscr{X} \longrightarrow B$ can also give $I(A;B) > 0$. At the level of two events, no causal asymmetry can be derived from statistical relations. Two events are either correlated or not, and correlation gives no clue as to which is cause and which is effect, or whether both are effects of some common cause.

6.3.2 Three and More Variables

At this point, it may seem that an elaborate notation has been introduced merely to state the obvious: events whose outcomes are correlated are causally connected, either by standing in a cause–effect relationship to each other or by possessing some common cause. The usefulness of introducing directed graphs to model causal situations arises when one desires to keep track of the causal relationships between more than two variables. In addition, with more than two variables, the asymmetric nature of the cause–effect relationship gives rise to recognizable statistical patterns.

The simplest causal asymmetry to have a statistical signature is the 'causal fork' (Reichenbach 1956). Consider two causes that have a common effect: $A \longrightarrow B \longleftarrow C$, for example, the exclusive OR gate of the previous section. In this causal model, the only connection between A and C is the fact that they have a common effect. But there is no reason why two events should be correlated simply because they have a common effect. Therefore, $A \longrightarrow B \longleftarrow C$ implies $I(A;C) = 0$. Now suppose that the outcome of B is fixed to 0. By the rule for an exclusive OR gate, if A is 0 then C must be 0, if A is 1 then C must be 1: fixing the outcome of B implies perfect correlation between the outcomes of A and of C. So $A \longrightarrow B \longleftarrow C$ implies that $I(A;C) = 0$, but $I(A;C|B)$ need not equal 0. Now consider the same model, but with the direction of the cause–effect relationships reversed: $A \longleftarrow B \longrightarrow C$. Here B is a common cause of both A and C. Since A and C have a common

cause, there is no reason to expect that $I(A;C) = 0$. In addition, since the only causal connection between A and C is their common cause B, fixing the value of B should destroy the correlation between A and C: $I(A;C|B) = 0$. That is, A and C are correlated because variation in the outcome of B induces a correlated variation in the outcomes of A and C. In the absence of any variation in B, there is no correlation between A and C. An example is a logic gate that has one input, B, and two outputs, $A = B$, and $C = 1 - B$; here A and C are perfectly anticorrelated, but fixing the value of B destroys that correlation.

The two causal models with the cause–effect relationships reversed imply entirely different statistical independence relations. Having a common effect does not induce correlation between events, while having a common cause does. Controlling for the outcome of a common effect can make the outcomes of its causes correlated, while controlling for the outcome of a common cause has the opposite result. Reichenbach identified the different independence relationships implied by common causes and by common effects as forming the basis for our identification of causal asymmetries in the world around us. This asymmetry in causation is responsible for the primary psychological arrow of time, our belief that we can change the future, but not the past. The difference between common cause and common effect implies that correlations between events in the present are to be ascribed to common causes in the past. In particular, the correlation between our memories of past actions, and those events that our actions have affected, lead us to identify our past actions as a common cause of those present events and present memories. However, there is no reason why a future choice of action should generate correlation between present events and present state of mind. Therefore, insofar as our choices are designed to effect a positive correlation between our desires and the state of affairs in the world around us, this correlation lies in the future; the past is beyond our control.

The general rule for deriving independence relations between events in causal models with many variables can now be presented. The following result is due to Pearl (Pearl 1988), and results from applying the two rules given above: 1) two events whose only causal connection is a common effect should not be correlated unless the outcome of their common effect is fixed; 2) a common cause tends to induce correlation between its effects unless its outcome is fixed. Suppose that a set of events $\mathbf{Y}$ have their outcomes fixed. Consider a path within the causal model. This path is capable of inducing correlation between the events that make up its endpoints if all the common effects along the path either belong to the set $\mathbf{Y}$ or have some descendant in $\mathbf{Y}$, and if no other events in the path belong to $\mathbf{Y}$. Such a path is called *open*. A path that is not open is called *closed*. If there is no open path between A and B given $\mathbf{Y}$, then $I(A;B|\mathbf{Y}) = 0$.

That is, a path is capable of generating correlation between its endpoints given $\mathbf{Y}$ if none of the common or intermediate causes — places where correlation is generated or propagated — along the path are fixed by fixing $\mathbf{Y}$, and if all the common effects — places where the propagation of correlation breaks down —

have been fixed to some degree, thus allowing correlation between their causes. So, for example, in the path $A \longrightarrow B \longleftarrow C \longrightarrow D \longleftarrow E$, fixing the outcomes of B and D allows correlation between A and E, while fixing the outcomes of C and D allows no correlation between A and E.

The general rule for deriving conditional independence relations from causal models can now be given. Consider a causal model, given by a directed graph whose vertices represent both the events about which statistics are available, and also 'hidden' events about which no statistical information is available. Consider three non–overlapping sets of events, **X**, **Y** and **Z**, taken from the events within the model about which statistics are available. **X** and **Z** are conditionally independent given **Y**, $I(\mathbf{X};\mathbf{Z}|\mathbf{Y}) = 0$, if there are no open paths between events $X \in \mathbf{X}$ and $Z \in \mathbf{Z}$ given **Y**. If the causal model correctly represents the cause–effect relationships between events, then the conditional independence relationships derived from the model must hold. Statistically significant deviations from the implied independence relations falsify the model.

In contrast, the existence of independence relationships above and beyond those implied by the causal model does not falsify the model. Two events may possess an open path between them, and still have statistically independent outcomes, as in the exclusive *OR* gate above. However, in the absence of detailed knowledge of the actual form of causal influence, one generally expects the conditional independence relationships implied by the causal model to be the only ones actually present in the data. For an open path not to generate correlation between its endpoints requires special sorts of causal influence along the path; in the exclusive *OR* gate example, variation in one of the inputs fails to generate a correlated variation in the output if the values 0 and 1 for the other input are equal. Any deviation from equiprobability on the part of the second input will allow correlation between the other input and the output. A 'generic' open path generates correlation between its endpoints.

6.4 Bayesian Networks

Given these methods for deriving conditional independence relations from causal models, one can ask, When do two causal models imply the same set of independence relationships? The answer to this question is particularly simple when one restricts one's attention to causal models whose graphs contain neither directed loops nor unobserved variables. Such models, represented by directed, acyclic graphs, are called Bayesian Networks (Pearl 1988).

The first point to note is that two directed, acyclic graphs that have different links (ignoring the directionality of those links) imply different conditional independence relations. Any two events that are not directly linked can be made independent by fixing the values of some set of events, while no conditional independence relation can be derived between two events that are directly linked. So two Bayesian

Networks that imply the same set of independence relations must have links in the same places.

The second point to note is that two Bayesian Networks with links in the same places, but with different unlinked common effects, imply different independence relations. An unlinked common effect is one such as $A \longrightarrow B \longleftarrow C$, where the two causes have no direct link between them. An acyclic model that contains $A \longrightarrow B \longleftarrow C$ always implies some conditional independence relation between A and C in which the value of B is *not* fixed. Any other set of directions for the links, $A \longrightarrow B \longrightarrow C$, $A \longleftarrow B \longleftarrow C$, or $A \longleftarrow B \longrightarrow C$, gives conditional independence between A and C *only* if B is fixed. So two Bayesian Networks that imply the same set of independence relations must have links in the same places, and the same set of unlinked common effects.

It is straightforward to verify that once the positions of the links and the unlinked common effects are given, the direction of the remainder of the links may be varied in any way that does not produce a directed loop or a new unlinked common effect — all Bayesian Networks obtained by such a process imply the same set of conditional independence relationships. So two Bayesian Networks imply the same set of conditional independence relations if and only if they share the same link locations and the same unlinked common effects. This result was derived by the author and used in analyzing financial data (1986–87); the same result was derived independently by Pearl and Verma (Pearl, Verma 1990). Similar results can be derived if the no directed loop and no unobserved event restrictions are relaxed.

6.5 Causal Asymmetry in Physical Systems

The notion of causality in physics is the same as that assumed as a basis for causal modelling: variation in the outcome of the cause produces a correlated variation in the outcome of the effect. The methods of modelling causal systems given here, unsurprisingly, give correct results when applied to physical systems. As an example, the requirement in electrodynamics that the source–free part of the incoming electromagnetic field vanish is a way of realizing the requirement that correlated variation between the motions of charged particles be caused by the motions of charged particles in the past. Some subtleties arise, however.

Since the causal models considered here correspond to directed graphs, to apply them to physical systems, one must either discretize the physical system, or look at causal models that contain hidden variables. In addition, causal models are Markovian in nature, defining causal relationships in terms of conditional probabilities, while nondissipative physical systems are generally characterized by Hamiltonian mechanics, a very particular type of Markov process. The deterministic nature of Hamiltonian mechanics implies the existence of statistical independence relations above and beyond those simply implied by the causal structure of a system. For example, controlling for the initial or final state of a Hamiltonian system completely

determines the system's trajectory, and destroys the statistical correlation between all variables, whether common causes or common effects. As a result, for Hamiltonian systems in the absence of noise, causal asymmetry cannot generally be derived from statistics. Whenever noise is introduced, however, as in Brownian motion, or in other systems described by master equations, application of the above methods generally results in a unique set of statistical independence relations. In such cases, statistical relations can be used completely to determine the causal structure of the system.

The methods described in this article are classical in nature, and do not take into account quantum mechanics. Quantum mechanical correlations can violate Bell's inequality: that is, the correlation between two spins in the Bohm version of the Einstein–Podolsky–Rosen *gedanken* experiment is of a form that cannot be reduced to zero by controlling for the value of a common cause (Bell 1964, Bohm 1951, Einstein, Podolsky, Rosen 1935). This result does not invalidate the present work, however. The correlation between the spins is still due to a common cause in the past: this common cause — the S–wave state out of which the spins arise — is fundamentally non–classical, and is not an 'event' whose outcomes can be assigned probabilities. Such quantum–mechanical causes can be included in causal models as hidden common causes. The resulting causal models, though expressed in terms of classical probabilities, are perfectly consistent both with quantum mechanics and with experiment. Bell's inequalities tell us only that such common causes cannot be resolved by experiment: hidden quantum–mechanical causes will always remain hidden.

6.6 Conclusion

The methods presented here are useful in ruling out causal models that predict independence relations that are not realized by the data. If hidden common causes can be ruled out *a priori*, then in many occasions, one and only one causal model is consistent with the data. If common causes cannot be ruled out, then although one can still rule out causal models that are inconsistent with the data, no unique model can be derived from the data; one can always postulate a model with many different hidden causes (e.g., a conspiracy theory), that explains the correlations in the data as accurately as a model with few or no hidden causes. In such cases a further principle, such as Occam's razor, must be introduced to identify the most plausible causal model.

In closing, it should be noted that when applied to the universe as a whole, the causal models presented here require a particular type of initial condition. If correlation in the present is to be ascribed to common causes in the past, then at the unique moment at which there was no past, there should be no correlation. Although one must be careful about extrapolating classical methods back to a quantum initial condition, the present work implies that in addition to being in a state of low entropy, the universe began in a state with no correlation between spacelike separated points

beyond that required by the Heisenberg uncertainty principle. In fact, Euclidean quantum gravity calculations point to just such an initial condition (Halliwell 1994, Laflamme 1994).

References

Bell, J.S. (1964) On the Einstein Podolsky Rosen paradox. *Physics*, **1**, 195–200.

Bohm, D. (1951) *Quantum Theory*, Prentice Hall, Englewood Cliffs.

Einstein, A., Podolsky, B., and Rosen, N. (1935) Can quantum–mechanical description of physical reality be considered complete? *Physical Review*, **47**, 777–780.

Halliwell, J.J. (1994) This volume.

Hume, D. (1739) *A Treatise of Human Nature*, John Noon, London. Reprinted 1906, Clarendon Press, Oxford.

Laflamme, R. (1994) This volume.

Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems*, Morgan Kauffman, San Mateo.

Pearl, J., and Verma, T.S. (1990) Equivalence and synthesis of causal models, Technical Report R-150, Cognitive Systems Laboratory, University of California, Los Angeles.

Reichenbach, H. (1956) *The Direction of Time*, University of California Press, Berkeley.

Russell, B. (1929) *Mysticism and Logic*, Norton, New York.

Shannon, C.E., and Weaver, W. (1949) *The Mathematical Theory of Communication*, University of Illinois Press, Urbana.