Elementary Lectures on Thirteen Problems for Science and Its Philosophy

These lectures are on problems I think are fundamental in philosophy of science. I have thought at various times about each of them, but never thought I knew all (or in some cases any) of the answers. The lectures are intended chiefly for students and scientists who want to think about questions that should arise in many scientific subjects, but are seldom discussed carefully in scientific texts or courses. The lectures do not try to provide all the answers, or even to survey the full range of answers others think they have found, although I will talk about many of those ideas. My principal intention is to make the issues as clear, vivid and as simple as I can. There is nothing sophisticated or even technical about the content, and I have avoided as much formalism as possible. Anyone who remembers a bit of high school algebra and some elementary probability—and not much of either—should, I very much hope, be able to understand the issues. Contributing to their solution may take a good deal more. Here are the titles of the problems:

1. Lecture 1: What is Probability? (Drafted)
2. Lecture 2: What is Causality? (Drafted)
3. Lecture 3: How Does the Big Arise from the Small? (Drafted)
4. Lecture 4: Is Explanation a Guide to Truth? (Drafted)
5. Lecture 5. Is There Any Logic to Discovery?
6. Lecture 6: Can Science Be Automated? (Drafted)
7. Lecture 7: How Can Scientific Revolutions Be Rational?
8. Lecture 8: How Can There Be Social Sciences?
9. Lecture 9: How Does the Brain Compute?
10. Lecture 10: How is Conscious Content Made?
11. Lecture 11: Can We Make an Android Baby?
12. Lecture 12: What Science Should Be Done?
13. Lecture 13: How Should Science Guide Politics and Morals?

Later lectures often presuppose some familiarity with the lectures that precede them.

Lecture 2: What Is Causality?

The short answer is that we don’t know. The long answer is very much like the answer for probability—there are various proposals, none of them entirely adequate, we learn something from them, and eventually we find a set of axioms that systematize a lot (but not all) of scientific practice. Just as some of the axioms about probability connect probability with causation, the interesting axioms about causality connect that notion with probability. Just as there are unresolved issues about the meaning of “probability,” there are unresolved issues about the meaning of “causality.”

Some philosophers, and many statisticians, think there is something especially unclear about the notion of causality and pretend to ignore it—generally, only by ignoring the word but not the idea. Terry Speed, a distinguished statistician, says that he recommends thinking and talking about causality as little as possible. That is flippancy: Speed testified about causality for the defense at one of the most famous murder trials of the 20th century, the trial of O.J. Simpson, accused of slashing the throat of his wife and of a young man who was at her house. In fact, Speed testified to a principle about causality we will discuss in this lecture. Refusing to think about a concept so imbedded in science only makes one—well, thoughtless.

The reasons for skepticism vary, but include the following:

- Judgements about causation are unstable—changes in background variables can change whether X is said to cause Y.

- Judgements about causation are often uncertain, even when we know all of the “facts” of the case—the literature in philosophy of science is filled with worrying examples.

- Judgements of causation imply claims about what did not happen, about what could or would have happened in various circumstances that were not the actual circumstances.

- Causality is not part of the physics of the world—there is no physical theory of what constitutes "causation."

All of these are true, but anyone who thinks these are sufficient reasons to be skeptical about causation but not about probability should do some further thinking: we have just seen that the notion of probability has all of these features and more.

Causation poses (at least) two different problems, one about individual cases of causal relations, the other about causal regularities. The first kind of problem deals with when, given other knowledge, we should say that one particular event causes another particular event, however the events are described, as when we say that Shlomo became ill because he smoked. The second kind of problem deals with causal relations between *kinds* of events, or between variables, as when we say that smoking causes illness. Problems of the first sort are dealt with almost exclusively in the philosophical literature; problems of the second sort have been addressed for the most part by statisticians and computer scientists. The division of interest is to some extent artificial. While there is some reason why statisticians should not worry about causation between particular events (*singular* causation), there is no reason why philosophers should not think about causal regularities. I will start with the second problem, where I think there is a tolerable theory. Not, to be sure, a theory that says what a causal regularity *is*, but a theory that says something useful about the connections between causal regularities, probabilities, and actions.

1. Elements of Experimental Inference

We discover and manipulate our everyday world with some simple principles relating causal connections with probability. Suppose I think I have discovered a wonderful new drug, HCP, that prevents hiccups. What do I have to do to convince the Food and Drug

Administration that HCP prevents hiccups? Well, first I have to make a showing that HCP is harmless. So I give varying doses of HCP, including no dose at all, to randomly selected groups of mice, rats, pigs, maybe monkeys, members of each species in each experimental group as genetically alike as possible, and keep track of their health and longevity. I calculate the life span and health measures within each group, and I compare the averages of different groups with a statistical test. Most commonly, the test is of the hypothesis that there is no difference in the average value of a measure between groups. If the differences are small enough, and the sample is large enough, my statistical tables (or nowadays, computer program), says that if the hypothesis is true the probability of a difference as large as I observed is pretty fair, and if various alternative hypotheses were true, I would probably have observed a bigger difference. So I report a *negative causal* conclusion to the FDA: HCP does not cause morbidity or death in any of these animals. (Actually, I am being optimistic, and so is the FDA. For all I know from my experiments, HCP might cause some animals to live longer and others to live shorter lives, some to be healthier than normal and others to be sicker, and all of these effects might just cancel out when I take the average over a group treated with HCP. That kind of difference would show up as greater variation in health or longevity within the HCP group than within the control group.)

Not enough for the FDA, of course; they want evidence that HCP does something good. So, I conduct trials with people, both with people who report no particular problem with hiccups and with people, like me, who hiccup a lot. I randomly assign them to groups that receive various doses of HCP (never exceeding the proportionate dosage by body weight of the largest dose I gave to animals) and to groups that do not I measure how many hiccupping events each person has over the course of a month, and how many hiccups there are in each hiccupping event. As before, I compute the averages for each group, and test the hypothesis that there is no difference between the groups. This time, however, the statistical tables say that the differences I observe are very improbable if the hypothesis is true. And this time I report a *positive causal* conclusion to the FDA, albeit a positive conclusion of a negative kind: HCP reduces hiccups.

What is the reasoning? A test of the hypothesis that there is no difference between group averages is a test of a particular consequence of a more general hypothesis: that the treatment (HCP or no HCP) is *independent* of the outcome (longevity or hiccupping). So is a test of whether the groups have the same measure of variation. Informally, independence means that the information that a group received or did not receive HCP provides no information about longevity or hiccupping in the group. If I were to assume that everyone had the same disposition to hiccup without treatment, I would say (on the *null* hypothesis) that the probability of any person hiccupping when treated with HCP is the same as the probability of hiccupping when not on HCP. I have used this connection between group (or, optimistically, individual) probabilities and the causal role of HCP in two ways: when I found independence (or at least no difference in averages or variances) for HCP treatment and animal health and longevity, I claimed there was no causal connection between HCP and health and longevity. When I found dependence, I claimed there was a causal connection. A picture will do:

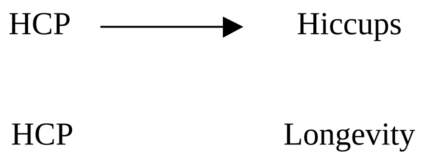


Figure 1

Why did I randomize the selection which animals and people were given various treatments, including no treatment? One reason is that I then knew the probabilities of various treatments, which helped in selecting a statistical test. The other reason is that randomization should reduce the chances that there is an association in the experiment between hiccupping and HCP that arises because the people in the HCP group were less likely to hiccup anyway. Why would that be? Well, suppose for example people who are more likely to hiccup volunteered first for the experiment, and I assigned the first arrivals to the control group that is not treated with HCP. Then a propensity to hiccup would actually be a cause of not being treated with HCP, and would also be a cause of hiccupping during the duration of the experiment.

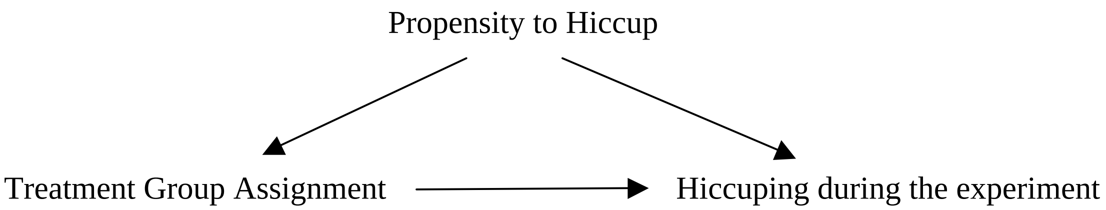


Figure 2

Any association I might find between HCP and hiccupping would be produced by two mechanisms: the influence, if any, of HCP on hiccupping and the joint influence of propensity to hiccup on treatment and on hiccupping, during the experiment, and there would be no way to separate how much of the association is due to which mechanism. (There is another principle here: Common causes of features produce associations of the features.)

Ramdomization doesn’t guarantee that there would be no such “confounding” factors producing an association between treatment and outcome, but it reduces the chances. Suppose I discovered that my assistant, a dull boy, had given preliminary treatment assignments based on the first-come/no-treatment rule. If I had carried out the experiment with this treatment assignment, the potential causal relations would be as in figure 2, but fortunately I reassigned people to treatment groups randomly. In doing so, I broke the causal connection between propensity to hiccup and treatment assignment:

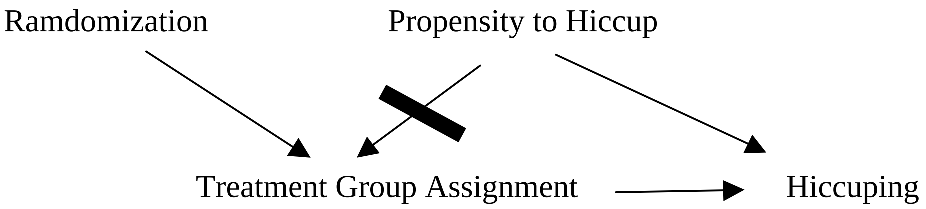


Figure 3

But by randomizing I did not alter the influence, if any, of HCP on hiccupping. With my reassignment I can’t be certain that there are no confounding associations, but they for sure cannot arise from common causes of treatment assignment and hiccupping.

Randomization is not the only trick in experimentation. Beliefs held by experimenters and by human subjects famously can influence outcomes of experiments. ESP experiments, for example, used to be conducted with a person counting whether the “sender” and “receiver” identified the same card in a pack of Zenner cards (cards with simple shapes on one side). It was found that counters who believed in ESP recorded more agreements than counters who were skeptical. If I knew which subjects received HCP I might give them, even unintentionally, special encouragement not to hiccup. If the subjects knew they received HCP, they might make special efforts not to hiccup. The causal relations would look something like this, even with randomization:

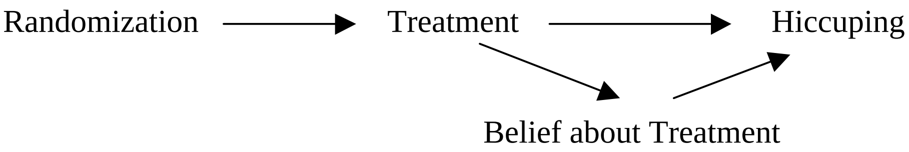


Figure 4

Once again, there would be two mechanisms potentially contributing to any association between the treatment group and hiccupping, and no way to isolate the specific effect of HCP. The problem is usually finessed by “blinding” the subjects and the experimenters, so that neither knows who received what treatment or non-treatment until all of the hiccupping results are in. We might blind subjects by giving the non-treatment group a placebo that looks, smells and tastes like HCP. Blinding effectively cuts off one causal pathway:

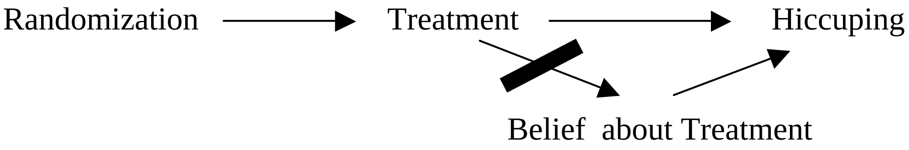


Figure 5

Finally, one more complication, using an example of Donald Rubin's (a statistician who just hates these diagrammatic representations of causal and probability relations.) Suppose the Ethics Committee of my university will not allow me to simply randomize treatments: they insist that I somehow measure the propensity of subjects to hiccup, and give those with high hiccupping propensity a larger chance of being assigned to the group to be treated with HCP. I can randomize enough so that there are some subjects of every propensity in all of the groups, but I must, as it were, make the treatment assignments with a biased coin. Now it looks as if I am back in the pickle of figure 2.

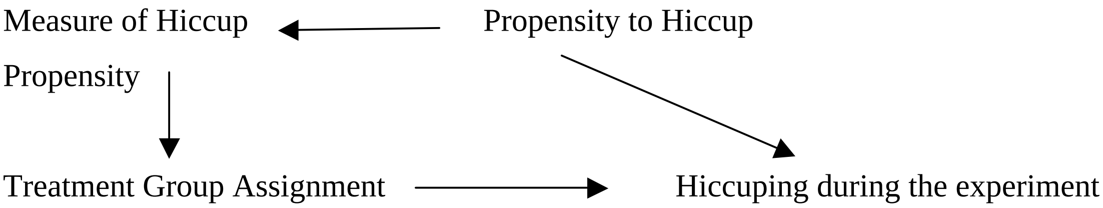

Measure of Hiccup Propensity
Propensity to Hiccup
Treatment Group Assignment
Hiccuping during the experiment

Figure 6

Once again, there are two mechanisms that can produce an association between HCP and hiccupping. How can I separate them? Rubin proposes I do this: instead of measuring the association between HCP treatment and hiccupping in all people in each group, measure that association within subsets of people in each group who have the same measure of hiccup propensity. More abstractly, rather than testing whether treatment and hiccupping are associated, I should test whether treatment and hiccupping are associated *conditional* on the measure of hiccup propensity.

The idea is that among people who are alike in this measure, propensity to hiccup cannot influence treatment assignment. Conditioning on measure of hiccup assignment works, statistically, much like an experimental intervention that fixes that measure: variations in propensity to hiccup cannot cause variations of the measure of propensity to hiccup within groups of people who are alike on that measure.

Everything I have said about discovering the effect of HCP on hiccupping is much more difficult if HCP isn't given experimentally. Suppose, for example, that HCP is a substance that occurs naturally in people, and whose concentrations in people can be measured. If I randomly select a sample of people and measure their HCP levels and their hiccupping and find an association, how do I know the association is due to the influence of HCP on hiccupping? Maybe hiccupping reduces HCP levels. Maybe something else in the body influences HCP levels and hiccupping. I'm stuck.

So much for hiccupping. How do these various principles and finesses fit together into a reasonably unified account of simple experimentation to discover causal relationships? Here is the story. Represent variables in an experiment or study as nodes or vertices. Put a directed edge—an arrow—from one variable to another if the first "directly" causes the second. Roughly, that means there is an arrow from X to Y if there is at least one causal pathway from X to Y for which no other variable represented in the study is in between X and Y on that pathway. The result is a directed graph, and I will assume the graph has no cycles—no sequence of directed edges or arrows in the same direction that start at a variable and end up at that same variable. Assume that there is a probability distribution assigning a probability to every arrangement of values of the represented variables. Not just any probability distribution will do, only a probability distribution that satisfies a special condition for whatever the directed graph may be: In stating the condition we need two further ideas, *X is independent of Y conditional on Z* if and only for every value of X, and every value of Y and every value of Z, the probability of the product of X and Y conditional on Z equals the probability of X conditional on Z multiplied by the probability of Y conditional on Z (provided the value of Z has non-zero probability). *A set of variables is causally sufficient* if and only if for every variable U that influences two or more variables in the set through causal paths that intersect only in U is also in the set of variables.

Here is the condition connecting any allowable probability distribution on the variables with the structure of the directed graph representing the causal relations among the variables (and incidentally, the condition to which Terry Speed testified at the O.J. Simpson trial—appropriately, since Speed was one of the first to formulate it):

Markov Assumption: In a causally sufficient set of variables with causal relations represented by a graph G, each variable is independent (in probability) of variables that are not its effects conditional on all of its direct causes in G.

If you think for a moment about the positive inferences described previously—when randomized experiment shows that HCP prevents hiccups, you will see that it is an instance, an example, of an application of the Markov assumption. The only variables are HCP, the hiccupping outcome, and the randomization, and because of randomization the system is causally sufficient and hiccupping in the experiment does not cause the treatment a subject receives. If HCP is not a cause (in the story, a preventive cause) of hiccupping, then by the Markov Assumption, HCP and hiccupping should be independent. But they are not independent, hence we conclude (or conjecture or surmise, depending on the sample size and no doubt on a lot of other considerations) that HCP is a cause of (not) hiccupping. Again, if you think about the Rubin example, it makes use of the Markov Assumption: the hiccupping measure is the only direct cause of treatment assignment, and therefore, by the Markov Assumption, treatment assignment is independent of propensity to hiccup conditional on the hiccupping measure.

Our examples used a second principle as well, when we inferred from the absence of an association between HCP treatment and longevity in animals that longevity is not influenced by HCP. The principle is a kind of converse of the Markov Assumption, another restriction on the connection between a causal graph and a probability distribution on the variables in the graph.

Faithfulness Assumption: Two variables in a causal graph are independent conditional on a set of other variables if and only if the Markov Assumption applied to the graph entails that independence.

So if you think about the story that went with Figure 1, I hope you will see that the treatment had no association with HCP, and we concluded that the true causal graph has no arrow connecting HCP and longevity—that is, we drew the conclusion that the correct causal graph implies, by the Markov condition, that HCP is independent of longevity.

We need one further definition and the fundamentals of the unified story (for these simple cases) will be complete. Suppose we have a set V of variables whose causal relations are described by a causal graph with an appropriate probability distribution P on the variables. Let X be a variable in V. We will say that a *policy variable for X* is a new variable, Ix, with no edges directed into it and a single edge out of it, directed into X. A policy variable must have at least two values. We further assume there is an expanded probability distribution P* that includes all of the variables in P and Ix as well and that satisfies the Markov Assumption for the graph that adds the Ix -> X connection, and that P* "fits with" P in the following way: for one of the values, say, k, of Ix, P* conditional on Ix = k is a probability distribution on V that is identical with P, while for any other value, say j, of Ix, the probability distribution on X is different from the distribution P imposes on X, but all of the conditional probability relations in P are unaltered, except for the conditional probabilities of X on other variables in V. An *intervention on X* is simply a change in value of the policy variable Ix for X, from k to j.

"Randomization" is a policy variable for the HCP variable in our examples. Without randomization, because of my dull assistant, the probability a subject would receive HCP depends on his hiccupping propensity; after randomization, the probability of receiving HCP is changed, but the probability of hiccupping conditional on receiving HCP is unchanged. (To think this way, we sometimes need to imagine some probabilities that are not revealed in actual frequencies—e.g., the probability that an individual in the subject pool before treatment who does not yet (or ever in the experiment) actually have HCP would hiccup conditional on receiving HCP.)

Suppose we know a causal structure, or think we do, or want to consider one as a hypothesis. If we are planning actions, we want to use our causal knowledge (or at least our beliefs) to predict outcomes of actions. If we are testing theories, we want to be able to predict the outcome of an experiment according to a hypothetical causal structure. The Markov and Faithfulness Assumptions, and the idea of policy variables, permit us to do so in many cases. First consider the case in which we know, or assume, a probability distribution and the causal relations among a causally sufficient set of variables. So we can apply the Markov Assumption to calculate the probability of values of any variable represented conditional on any values of any other represented variables. If we are intervening to fix a value for a variable X, or to randomize X, we need only break any arrows directed into X, impose the intervention value or probability distribution on X, and leave the conditional probabilities of all other variables unchanged.

When the variables for which we have probabilities form a causally insufficient subset of the represented variables—i.e., we know, or think we know, more about the causal relations than about the probabilities—things are more complicated. There are, however, algorithms for determining in such cases whether the probability of a variable can be computed when there is an intervention on another variable, and for computing the actual probability values. For example, suppose we were considering the following hypothesis,

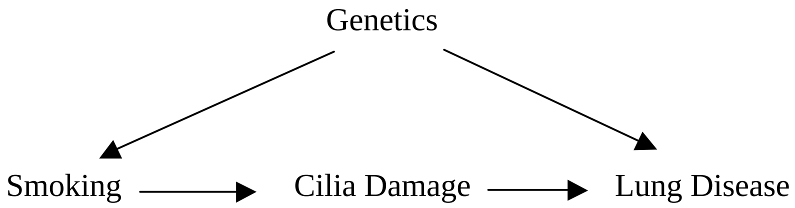


Figure 8

where genetics is unmeasured, but we have observed the joint frequency and estimated the probability of Smoking, We want to predict the effect on Lung Disease of Cilia Damage and Smoking. Suppose we have a drug that inhibits Cilia damage and a smoking cessation program. We would like to use the hypothesis to predict something about each of these treatments. We could use statistical procedures to try to estimate the value of genetics for each person, and then apply our general theory for causally sufficient systems to but that is not very promising—there are roughly 30,000 different human genes, and that makes for a lot of different values for the Genetics variable. Instead, we can use the causal hypotheses to predict the effect on Lung Disease of stopping Cilia Damage—as in the example from Rubin, we need only condition the observed probably of Smoking. More surprisingly, the probability of Lung Disease on an intervention that stops smoking can also be predicted, even though Smoking and Lung Disease are directly confounded by Genetics.

There are many open problems about predicting with incomplete causal knowledge and incomplete probabilities. For example, in many causes a causal hypothesis may be incomplete. It might say, for example that two variables, X and Y, are associated, but not say whether X causes Y, or Y causes X, or something else unobserved causes both. We could of course try to calculate predictions of some other variable, Z, after an intervention on X using each of the alternative causal hypotheses, and see what range of predictions we get. But depending on how complex our theory is, there might be a great many alternatives. A correct, complete way to do the prediction directly in such cases is not yet known.

The procedures I have been discussing apply to any system in which the Markov and Faithfulness Assumptions hold, no matter what kind of probability distribution we are considering. But there are special prediction possibilities that hold for restricted families of probability distributions. For example, suppose we are dealing with Normally distributed variables, and the causal structure is of the form:

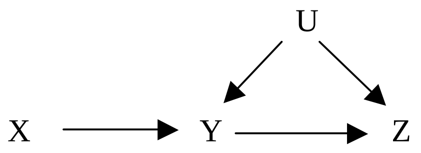


Figure 9

Suppose we have observed frequencies and estimated joint probabilities only for X, Y and Z. In this case the general procedures discussed above do not allow the prediction of the effect on Z of an intervention on Y. Nonetheless, it can be predicted, and the technique, known as instrumental variables, has long been used in economics. Less well known is that the same sort of thing is possible for certain other families of probability distributions. Suppose for example that Y occurs with some probability if X occurs or with some independent probability if U occurs, and similarly for Z with respect to Y and U. Then, again, the effect on Z of an intervention on U can be predicted from the probabilities before the intervention.

The general problem of what predictions are possible from what causal assumptions with what observed probabilities from what families of probability distributions, is still very much unsettled.

Our understanding of causal relations, many of our methods of discovering them from experiments and observations, and our methods for predicting from them, are *local.* Our descriptions of causal relations separate the world into two parts—the variables and relations being described or studied in a period and region, and everything else. Our claims about what does and doesn't cause what, and how strongly, presuppose that the rest of the world stays pretty much as it is. A big dose of cynanide will kill you, *but* not if something else intervenes to kill you first; a cobra bite will kill you, *but* not if you receive sufficient anti-venom in time. The price of rice in Tibet has no effect on who is chosen U.S. Secretary of State, *but* things could be rearranged so that one depends on the other. The planets will move in their orbits, *unless* something knocks them out of orbit. We have a principle for connecting a system we describe in causal terms with the rest of the world, and that is the idea of an intervention. Of course outside influences on a system need not be so neat as the idea of an intervention sketched above suggests—for example, whether intended or not, an intervention might directly affect several variables in a system. And nothing in the unified theory sketched above guarantees that any interventions are possible for a particular variable in a particular causal system.

With all of these limitations, an amazing variety of statistical methods for causal inference, and predictions about the effects of experimental interventions assuming causal hypotheses, are accounted for by these two axioms and the definition of an intervention. This is how we discover a lot of what we know about the world, and how we predict with what we know.

Let’s consider a prediction, in a famous (among statisticians) imaginary case that shows that for correct prediction we need both probability and causal knowledge. Suppose we have the following data for variable Y, with values y1 and y2, and variable X with values x1 and x2, for two groups, one with value z1 for variable Z and the other with value z2 for Z:

z1	y1	y2	Total
x1	18	12	30
x2	7	3	10
Total	25	15	40

z2	y1	y2	Total
x1	2	8	10
x2	9	21	30
Total	11	29	40

A little examination reveals that none of the variables are independent of any others. The frequency of y1 versus y2 varies with the value of X and the value of Z, and the frequency of the value of X likewise varies with the value of Z. Notice also, something odd about the conditional probabilities: $p(y1|x2, z1) > p(y1|x1, z1)$, and $p(y1|x2, z2) > p(y1|x1, z2)$, but $p(y1 | x2) < p(y1|x1)$. No matter which value of Z we condition on, also conditioning on x2 gives y1 a higher probability than does also conditioning on x1. But if we do not condition on any value of Z, conditioning on x1 gives y1 a higher probability than does conditioning on x2. No pair of the variables are independent conditional on the third.

Now consider two different ways these data could have been generated:

1. A medical experiment. X = treatment; x1 = treated; x2 = not treated; Y = Outcome; y1 = recovered; y2 = did not recover; Z = sex; z1 = male; z2 = female.

2. An agricultural experiment. X = variety of plant; x1 = white, x2 = black; Y = Yield; y1 = high, y2 = low; Z = height of plant; z1 = tall; z2 = short.

Here are the questions: If you want to produce the best medical effect, should the treatment be given, and to whom? If you want to produce the best yield, what variety of plant should be planted? The example is due to Lindley and Novick, two distinguished Bayesian statisticians. They give the following answers. For the medical case, no treatment should be given, i.e, x2 is the “non-treatment” of choice, because it has better recovery probabilities for males and better recovery probabilities for females. In the agricultural experiment, white plants should be grown (i.e., x1) because it has a better probability of high yield overall.

This may seem strange. After all, the probabilities are the same—the very same numbers. Why such different recommendations about interventions? Why indeed? I will not give Lindley and Novick’s explanation, but a different one (although the explanations are related). My answer is that the two different stories, one medical and the other agricultural, naturally lead us to different causal interpretations of the data, and the different causal interpretations suggest different effects of interventions.

Consider the medical case first. *Our task is to estimate the relative effects of treatment versus no treatment in the experiment, and then use that information to recommend how the general population should be treated, or not treated.* We assume that in the experiment someone’s sex is not caused by the medical treatment. Sex and Treatment, Z and X, are not independent in the data—more men received the treatment than did women. The dependency must then either be due to chance in randomly choosing subjects, or due to the influence of sex on which patients were treated and which were untreated. The causal structure in the experiment is:

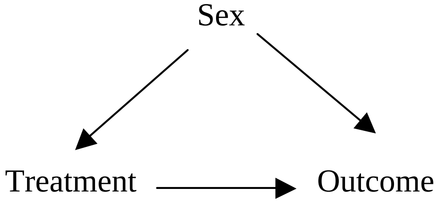


Figure 10

or

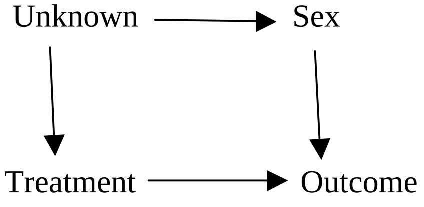


Figure 11

In either case know from the Markov Assumption how to compute the probability of recovery, y1, from an intervention,—a forced value of x1 or of x2--in a system with this description. We must eliminate the association between treatment choice and outcome due to the common cause (sex in figure 10; unknown in figure 11), and use the remaining association as our estimate of the effect of treatment choices on recovery. We can do that by conditioning on Z, on the sex of the subjects. We compute the probability of y1 conditional on x1, for example, and conditional on Z. The probability of y1 conditional on x1, or respectively on x2, is of course different for different values of Z, for males or females, but x2 is better in both cases.

In the agricultural interpretation of the experiment, heights of plants are correlated with variety of plant. That is most plausibly because the plant's variety influences its height, or something else—genetics, say—influences both the variety and the height of the plant. The height of the plant doesn't influence the variety.

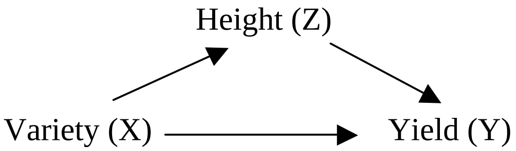


Figure 12

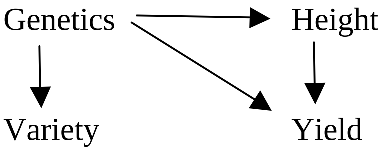
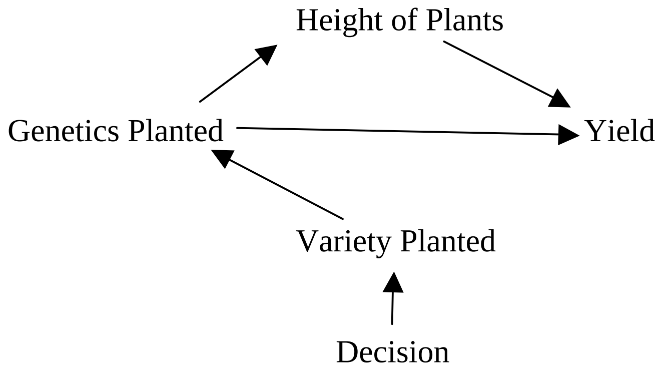


Figure 13

In recommending a planting policy, we know the variety of seed to be planted, black or white, but we do not know when we plant whether the plant will be short or tall. If in fact the plant variety influences height, as in Figure 9, then the variety influences the yield through two mechanisms, one direct, and the other through height. In that case, to assess the influence of variety on yield, we *should not* condition on height. To do so would be to discount one of the paths by which variety influences yield.

If we think of genetics as a common cause of variety and height, as in Figure 13, the issue is more subtle. In that case, plant variety is not a cause of yield at all. From the experiment we cannot measure the causal influence of variety on yield, because there is none. All we can do is measure the total association between variety and yield, and the associations conditional on the two values of height. Our problem is how we can use that information to decide whether we will get a better yield from planting white or black plants. In Figure 10, plant variety and genetics are *necessarily related*: genetics uniquely determines variety. That means that variety cannot be manipulated without manipulating genetics. There is no sense in which two plants have the same genetics but are different varieties. (We shouldn't assume genetics uniquely determine height—environmental factors obviously influence that variable.). In contrast, variety does not uniquely determine genetics, of course—two plants can be of the same variety but differ in some of their genes, for example one can have genes that tend to make it grow taller. But variety is associated with genetics—planting different varieties will result in a collection plants with different distributions of genes.

The variety of plant does not cause genetics *, but the intervention to plant a particular variety does cause plants with particular genetic features to be planted,* which causes heights of planted plants, which causes yield.



`

Figure 14

The distinctions, like causality generally, are subtle. The genetics of the plant causes its color, but doesn't cause the plant to be planted. In deciding what to plant, we will determine which plants are planted, and in doing so we influence the genetic features of what plants occupy a particular plot of land; the genetics in turn will influence the height of what grows in the plot, and the yield of that plot. The yield of a field is the sum of the yields of each plot of land. The critical thing is that the association between genetics and variety in the experiment in figure 13 is *must be the same* as the association between variety planted and genetics of what is planted in the policy we carry out in figure 14, so the causal influence of a decision to plant one variety rather than another is measured by the association in the experiment between variety and yield. The association measures are from the experiment, but the relevant causal relations in carrying out a planting policy have different causal structure than in the experiment, the causal structure of Figure 14 rather than figure 13. In figure 14, conditioning on height would be the wrong thing to do because it would discount a pathway from the intervention to the outcome.

Sounds ok, but not so fast. Suppose we were to consider a general practice in which particular doses of a drug of the kind used in the medical experiment are given. Couldn’t we argue as follows about new patients: We (or the physician) decide which patients, if any, get each particular drug dose (0 or whatever the treatment is in the experiment). Since we know the sex of the patients, we decide, for each dose, whether it goes to a male, to a female or to no one. But the sex of the recipient influences the probability of recovery.

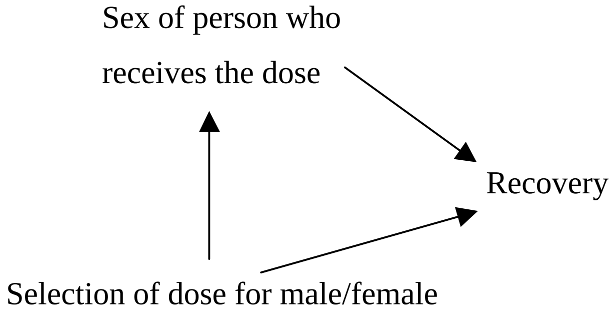

Sex of person who receives the dose
Recovery
Selection of dose for male/female

Figure 15

Hence there is a causal pathway from treatment decision to recovery through the sex of who receives the treatment, and hence, the arguent goes, just as in the agricultural case with figures 13 and 14, we *should not* condition on sex (Z) in assessing the effect of treatment. Hence, applying the probabilities in the experiment, we should give the treatment to everyone.

This argument only works, however, if in our policy of assigning doses to subjects taking account of their sex, the associations that result between dose and sex of the recipient are the same as in the experiment. In the practice proposed, sex and treatment are perfectly correlated, but they are not perfectly correlated in the experiment. Further our problem was to figure out which treatment was best as a policy for everyone. And if we were to give the same treatment to everyone, there would be no association between treatment and sex in the practice, and so once more the association of treatment and sex would not be the same in our practice as in the experiment. The argument fails.

The other way to think about the alternative practice is that the patient's sex influences our knowledge of the person's sex which influences the treatment the person is given (but not his or her sex), and we are back to the case of figure 12.

There are other problems about our discussion of causal regularities. Consider a bicycle. Turning the pedal turns the chain which turns the rear wheel. But on some bicycles, turning the rear wheel also turns the chain which turns the pedals. How can we represent this causal system? *Potentially*, the causal relations go in both directions, pedal to chain to wheel, and wheel to chain to pedal. In actual cases, there is an intervention that directs the causal relations one way or the other, not both. Or consider two meshed gears, arranged so that if either one turns, it turns the other. Suppose we intervene to spin each gear. If we spin the gears in complementary directions, each one spins and also spins the other:

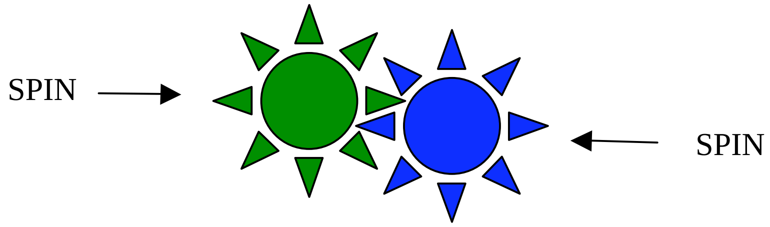


Figure 16

The causal diagram looks like this:

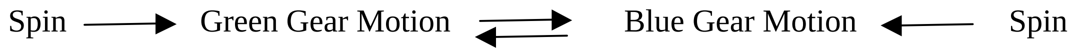


Figure 17

It turns out that in systems like this—cyclic systems, or feedback systems--the Markov Assumption is false. Structures like this also arise in economic theory, so the problem applies to science as well as simple mechanics. We could try to break the results of spinning both gears into discrete steps—first the green gear moves a little, which moves the blue gear a little, which then moves the green gear a little, and so on. A sequence like that is called a time series, and is in agreement with the Markov Assumption, but in the gear case the motions are continuous and essentially simultaneous. Fortunately, there is a generatlization of the Markov Assumption, called d-separation, that applies to causal diagrams with cycles as well as to those without cycles, and gives the same results as the Markov Assumption for causal graphs without cycles. The d-separation condition necessarily holds for *linear* systems, cyclic or not, but not for all non-linear systems.

There is still another problem. What if we spin the gears in opposite directions with equal amount of force, or what if push the pedals of the bicycle to make the wheel move in one direction, and push the wheel in the opposite direction at the same time with equal amount of force? Nothing moves. There is all sorts of causation going on—forces being applied—but nothing changes about the motions. We will consider problems like this again at the end of Lecture 3.

Lecture 3: How Does the Big Arise from the Small?

The great philosophical problem of ancient Greece was The One and the Many. I never understood quite what it was about, but our great philosophical problem, ever since the 17th century, is about how big things and their properties and relations can arise from itty bitty invisible things and their properties and relations. By the Big I mean middle sized dry and wet goods, the stuff of everyday life, not the galaxies and the cosmos. You can see the problem already in 17th century speculations about phenomena produced with air pumps. The great question then—one of them anyway--was what accounts for the fact that air resists compression. Nicholas Lemery proposed the air is like a tennis net that springs back when pressed; Robert Boyle proposed that air is made of particles shaped like little springs, like wood shavings. Our problems are harder. They arise from the apparent inconsistency with the quantum theory of principles that are used everyday in scientific experimentation and in ordinary life, and they arise in different forms without quantum theory at all. Understanding them requires review of some elementary aspects of how we discover and describe causal relations with middle sized dry and wet goods.

. So what is the problem?

1.1 Quantum Systems and Causal Systems: 3 Problems

Quantum theory poses at least three problems connecting the large and the small. I think two of them can be fairly straightforwardly resolved, although they often have been promoted as profound puzzles, and have prompted a variety of metaphysical fantasies. (It is fair, I think, to say, that some physicists cannot resist promoting their subject as full of paradoxes, and write philosophy of physics rather in the way Jean-Paul Sartre wrote about Being and Nothingness.)

1.1.1 Measurement and Intervention

Earlier, introducing the notion of a policy variable and the definition of interventions, I assimilated interventions to particular kinds of causal relations, which allows a uniform extension of the Markov Assumption, and a kind of stability or invariance: an intervention changes the probability of the variable intervened on, but leaves the conditional probabilities of all other variables unchanged. The Markov Assumption extends smoothly from the causal system under study to include interventions on that causal system. A mere observation of a variable in a system is a trivial intervention, one that changes nothing about the causal and probabilistic relations internal to the system. We can, without contortions, expand the system to include the observer and the causal relations to include the observation. . Who knows whether there are in the world singularly interventions and observations that are so smooth and so stable, but that is how we predict from what we think we know about causal relations in the Large, and while we are often surprised, we get along pretty well assuming our Big world consists of stable, consistently extendible systems.

One of Von Neumann's many great insights is that the quantum theory describes systems that are only extendible in a different way. A quantum system in a macroscopic environment—for example, an electron or photon in a box—has a state. The state of the particle determines a probability distribution for a collection of variables. For some of the variables the state entails a unique value, for other variables only a probability distribution. The state itself changes with time according to a solution to a partial differential equation, the Schrodinger Equation. There is one oddity: while all of the variables have a probability (counting 0 and 1 as probabilities), and various sets of variables have *joint* probabilities, the theory specfies no joint probability for observed values of all variables at once. It does specify various *conditional* probabilities: the probability that a observation that a system has property A given an observation that it has property B. The result is a bit odd: the theory gives us no probabilities for observing any specific values of all variables at once, but it does give us a probability for observing any value of any variable conditional on having first observed a set of values for any other Alt-collection of variables that do have a joint probability.

When an observation is made on the variables whose values are uniquely determined by the particle state, nothing changes except according to the differential equation. The probabilities for all other variables are as they were before, or as they should be according to the change of state according to the differential equation. Observation in these cases is just like observations of the Big. For variables for which the state of the system only determines a (not zero or one) probability, an observation finds some value or other, but the state of the system doesn't determine which value.. Suppose the system is in state S, and X is a variable that has a unique value x according to S, and Y is a variable for whose values S only determines various probabilities. When Y is observed, the probability that an observation of X will be x is no longer 1. So the observation has changed the state of the particle from S to something else, call it S,* and that change is not predicted by the Schrodinger equation when applied to the quantum system alone, but by another independent rule, called by Von Neumann the Projection Postulate, whose details need not concern us here.. Such an observation in the quantum world is roughly analogous to an intervention in the Big world: if W causes Y and Z the probability of Z conditional on Y is in general not the same as the probability of Z conditional on an intervention that forces a new value on Y.

If a quantum explanation is expanded to include not only a quantum system but also the process that measures a value of a variable in the system, the same principle applies: the observations are not predicted by the Schrodinger equation but by the combination of that equation and a rule for observations, i.e., for interventions of a special kind. So there is no getting away from this feature of the quantum theory.

There is a phenomenon about the large and the small here, but it is not exactly a problem or difficulty: the quantum theory is a local theory, a theory about dynamical systems in a stable environment, coupled with a theory about measurements as interventions in such systems. Problems about measurement only become unsolvable if we think of the theory as a theory of everything all at once, for all time, because then there is no way to get outside of the system and apply the Projection Postulate.. That conception of the theory prompts bizarre metaphysical speculations of the kind my former colleague, Eugen Wigner, proposed, that claim nothing ever happens until a human becomes aware that it happens (apparently, Wigner thought consciousness isn't inside the system), . That kind of talk may be fun, but it is not serious. We do have a theory of everything, everywhere, for all time, the general theory of relativity, but it is not a theory about every *aspect* of everything, only about gravitation. The productive combinations of quantum theory and the theory of gravitation have been local theories—theories about the behavior of particular systems in particular environments, even if they are Big things in Big environments. Attempts to unify general relativity and quantum theory in a general way have not proved fruitful, and the measurement phenomena of quantum theory suggest they should not be.

1.1.2 The Existence Problem

I remarked in passing that there is no joint probability distribution for observable values of variables in a quantum system. What that means bears thinking about. We have a bunch of variables, and a state of a quantum system that assigns each value, x, of each variable, X, a probability that a measurement of X will yield the value x. Similarly for the probability the Y = y. But the theory does not assign a probability that observations will yield A = a *and* B = b *and*….X = x *and* Y = y *and* Z = z, and so on, for all variables, or even for various subsets of all variables. Moreover, the probabilities assigned by quantum states cannot be expanded to a probability distribution that does make such assignments of probabilities to joint distributions.

What that means is that the probabilities assigned to observable quantities by a quantum state cannot be understood as frequencies of properties in a collection of objects, each object having a definite array of values for *all* of the observable quantities of the theory. In other words, we cannot even hypothetically describe a population of objects and their positions, momenta, angular momenta, energy, whatever, so that the joint frequencies of their values for all the variables are consistent with the joint probabilities that the quantum theory allows for various subsets of the variables.

One response, perhaps the most popular, is that the phenomenon shows that Alt-atomic particles *do not have* all of their observable properties simultaneously. That creates a great mystery, since, things in the Big World, which are made up of quantum things, do seem to have all of their properties simultaneously. There are two Alt-solutions: (1) what seems to be the case is not-- things don’t really have both positions and momenta and such simultaneously; and (2) things have interval valued positions and momenta and such simultaneously, but not precise values for them. The second idea needs a small comment.

Our physical theories use continuous variables, with an uncountable infinity of values. All of our numerical measurements turn out to be numbers that can be expressed as fractions. Perhaps all the others, and perhaps even many of the fractions, are just mathematical conveniences, corresponding to nothing physically real. A thing, then, doesn't have a position corresponding to an exact numerical description of its boundaries or its center of mass, it has only an interval of positions. And similarly for other variables. This idea can be reconciled with the quantum theory, because although there are no joint probabilities for position and momentum there are joint probabilities for *intervals* of position and momentum. The interval-valued solution has a potential problem that may reduce it to the "properties pop into existence when observed solution." I do not know how precisely the position or momentum of anything has ever been measured. But any measurement of position (of anything at any time) within interval e would, on the interval proposal, imply that positions that precise exist, and any measurement (on anything at any time) of momentum within interval d would, on the same proposal, imply that momentum values of that precision exist. If these two values are small enough, the quantum theory would again not assign joint probabilities and we could not conclude that one and the same thing has both properties with both precisions at the same time. Small interval values would have to pop into and out of existence.

The other response, which is simpler and less peculiar and for that reason more sensible, goes like this. The probabilities quantum theory postulates are about observations, and the quantum theory treats many observations as interventions that change the state of the system. Quantum systems do have precise values for all of their quantities, but these values are not in general the values that show up on observation. The observed values are the interactive result of the state of the system at the moment of the measurement intervention, the intervention itself and (perhaps) the singular value of the quantity at the moment of the measurement intervention. The quantum theory doesn't predict these observed values from the state of the system, only their frequency distribution, or probability, and it doesn't give any account of the role of the singular values in producing observed values. Nonetheless, they exist.

Hypotheses of this kind are known as “hidden variable theories,” for obvious reasons. Hidden variable theories are unwelcome in physics, for reasons that have to do with scientific methodology: with enough hidden variables, you can concoct myriad fantastic explanations of all kinds of phenomena. Physics makes progress by refusing to consider all those alternative theories. But when the choice is between a useful methodological principle, on the one hand, and making the connection between the large and the small mysterious or nonsensical, on the other, perhaps wisdom lies with hidden variables.

1.1.3 The Locality/Non-Locality Problem

I call this problem “non-locality” because that is the name it is traditionally given in the philosophical and physical literature, but I view it—and of course I think this is the correct way to view it—as a problem about the Markov Assumption. It is a real problem and it is this:

Experiments tell us that the behavior of microscopic obects—photons—are inconsistent with the assumptions we make about causal relations among Big things, and, worse, we must use these very assumptions in designing the experiments and reasoning about them. The quantum theory, predicts the experimental results.

The small is not only inconsistent with principles we use in discovery about the large, but even with how we make discoveries about the small. Wittgenstein said his first book was a ladder to be kicked away when one had understood it; our ladder comes back and hits us on the head.

The details of the design and conduct of the first experiment, known as the Aspect experiment, showing the bizarre behavior of the microcosm also show lots of uses of the Markov and Faithfulness assumptions in the arrangement of instruments and equipment, but I will pass on those details and consider instead a very simple idealization of the phenomenon, due to N. David Mermin.

Consider two detectors I and II that are spatially separated. Each detector has three settings, S = 1, 2 or 3. Further each detector has a red bulb R and a green bulb G. Pairs of particles are emitted from a source and enter the two detectors. There is no other physical connection of any kind we know of between the detectors.

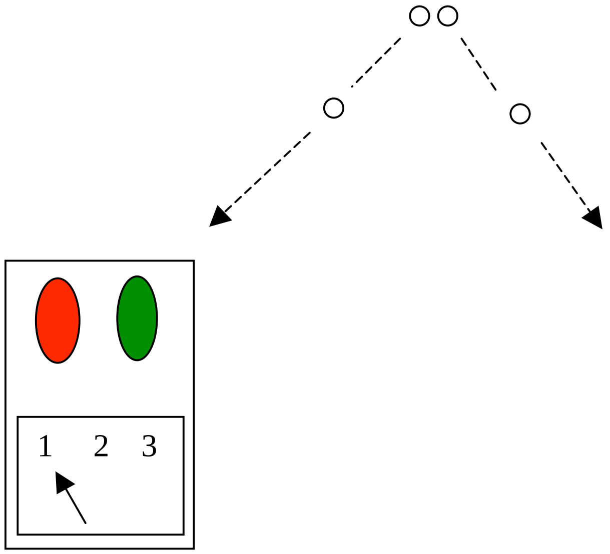


Figure 7

The detectors behave this way:(1) when both detectors are set to same value, no matter which, they both show red or they both show green. Red and green occur with equal frequency. (2) when the two detectors are set to any two *different* values, they show the same color, both red or both green, _ of the time—again, red and green occur with equal frequency in this case, and different colors 3/4 of the time—each combination of colors (I green, II red; I red, II green) equally often. We can show the whole story about the probabilities with a tedious but clear table

Table 1

Left Indicator Setting	Left Indicator Light Color	Right Indicator Setting	Right Indicator Light Color	Probability of the Two Light Colors Given the Settings	

1	Red	1	Red	1	
1	Green	1	Green	1	
2	Red	2	Red	1	
2	Green	2	Green	1	
3	Red	3	Red	1	
3	Green	3	Green	1	
1	Red	2	Red	1/8	
1	Green	2	Green	1/8	
1	Red	2	Green	3/8	
1	Green	2	Red	3/8	
2	Red	1	Red	1/8	
2	Green	1	Green	1/8	
2	Red	1	Green	3/8	
2	Green	1	Red	3/8	
1	Red	3	Red	1/8	
1	Green	3	Green	1/8	
1	Red	3	Green	3/8	
1	Green	3	Red	3/8	
3	Red	1	Red	1/8	
3	Green	1	Green	1/8	
3	Red	1	Green	3/8	
3	Green	1	Red	3/8	
2	Red	3	Red	1/8	
2	Green	3	Green	1/8	
2	Red	3	Green	3/8	
2	Green	3	Red	3/8	
3	Red	2	Red	1/8	
3	Green	2	Green	1/8	
3	Red	2	Green	3/8	
3	Green	2	Red	3/8	

The thing to notice immediately is that, no matter how we set the two detectors, the colors the detectors show will not be independent in probability. If both detectors are set at the same value, the probability that Detector II is red is 1 conditional on Detector I being red, and vice versa. If both detectors are set at different values, the probability that Detector II is green given that Detector I is red is three times the probability, on that same condition, that Detector II is red. Notice further, that someone at Detector I cannot use his settings of the detector to send signals or communications to someone at Detector II via the color that shows up at Detector II. For despite the fact that no matter how the detectors are set, the colors are correlated, the color at Detector II is independent in probability of the setting at Detector I.

Merwin puts the problem this way. The only explanation (he says) for the first six rows of the probability table is that the particles each have internal states that specify their response to each state of a detector. The internal states of each particle specify what color it will activate for each of the three settings of the detector. Since there are 2 possible colors for each detector setting, and three settings, there are 8 possible internal states for each particle. If and only if (Merwin says) both particles have the same internal states will the colors of the two detectors agree when they have the same setting, for all 3 possible settings. So the states of the particles have to be perfectly correlated, the same. If one particle will make a detector go red on setting 1, red on setting 2, and green on setting 3, so will the other. So the question becomes: *is there a probability distribution over these possible internal states of the two particles that, consistent with their perfect correlation, agrees with probability table?* There is not. In particular, there is no way to assign probabilities to the particle states so that when the settings of the detectors are different, the detector colors agree less than 1/3 of the time. Let's do another table. The columns indicate the settings of the two detectors when they are different, and the entries indicate for each state and pair of settings whether the colors of the detectors are the same or different.

Table 2

State	1, 2	2,1	1,3	3,1	2.3	3.2
RRR	Same	Same	Same	Same	Same	Same
RRG	Same	Same	Differ	Differ	Differ	Differ
RGR	Differ	Differ	Same	Same	Differ	Differ
GRR	Differ	Differ	Differ	Differ	Same	Same
RGG	Differ	Differ	Differ	Differ	Same	Same
GRG	Differ	Differ	Same	Same	Differ	Differ
GGR	Same	Same	Differ	Differ	Differ	Differ
GGG	Same	Same	Same	Same	Same	Same

In each row the fraction of cases in which the colors are the same is 1/3 or more. No matter what the relative frequency of the various particle states may be, if the detectors are set at any pair of distinct settings, the colors must be the same at least 1/3 of the time, but in the data for the experiment, for such settings the colors are the same only _ of the time.

So what does this have to do with Markov Assumption and so forth? Two things. On the one hand, the conclusion of the example, while not inconsistent with the Markov Assumption, is inconsistent with the conjunction of the Markov Assumption and the claim that the state of the particle is the only causal connection between the detectors. On the other hand, while Mirmin's reasoning is perfectly correct, his argument depends on using the Markov Assumption. I will represent Mirmin's account of his experiment as a causal graph, like this

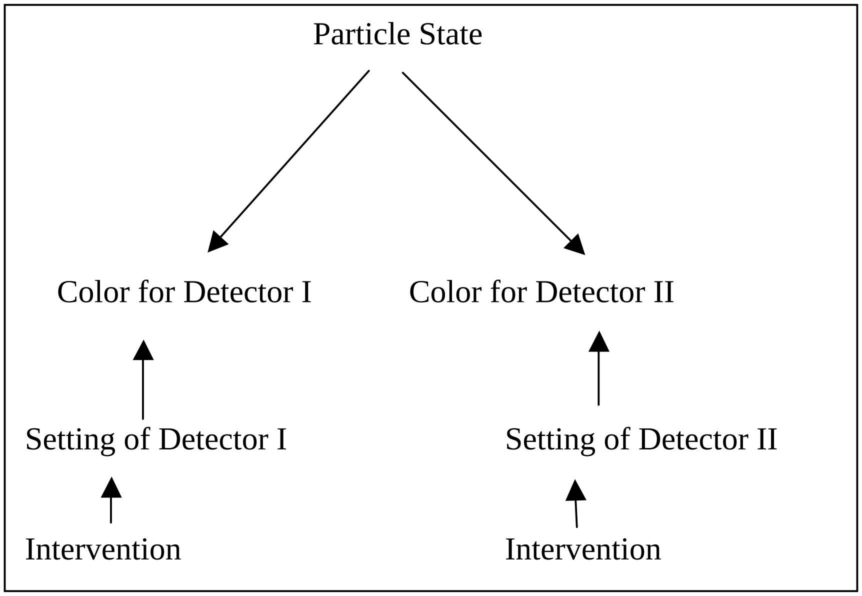


Figure 8

The causal diagram and the Markov Assumption explain why the setting of Detector I cannot be used to send a signal to Detector II via the color that appears at Detector II—there is no causal pathway from Setting of Detector I to Color for Detector II, or vice-versa, so the two variables must be independent. And the causal diagram explains why the colors at the two detectors are correlated: they have a common cause. Nonetheless, there is something very wrong. There is no causal pathway from Color for Detector I to Color for Detector II, or in the other direction. There is no common cause of detector colors other than Particle State. Since Color for Detector II is not an effect of Color for Detector I, and vice versa, the Markov Assumption says they if the causal graph above is correct, the detector colors should be independent of one another *conditional* on Particle State. Indeed, that is exactly what Mermin's particle states do imply. For example, given that the particle state is RRR, then Detector I is red and Detector II is red: no matter the settings and neither detector provides any information about the other detector not already entailed by the particle state. If the particle state is RGR, then no matter how Detector I is set, the color in Detector I gives no further information about color that will appear at Detector II. (The setting chosen for Detector II provides further information about the color that will show up for Detector II when the particle is in the

RGR state, but that is beside the point.). But Mirmin's argument shows that these particle states cannot be made consistent with the assumed observed frequencies of colors in each combination of settings shown in Table I. So there are logically just three alternatives: (1) Mirmin has sneaked in some extra assumption somewhere, or (2) the Markov Assumption is false for this case, or (3) there is no causal explanation of the correlations of the detector colors. Perhaps more than one of these alternatives is true.

Mirmin has certainly sneaked in some assumptions—all of them instances of the Markov Assumption--and the fact that he does not make them explicit may indicate that the Markov Assumption is so fundamental to our reasoning about experiments that we use it automatically, without notice. For *there is* a common cause explanation of the probabilities in Table 1! Here is the idea, first noted by Suppes and Zanotti in a more general case: Change the particle states so that they no longer just specify a color for each of the three settings of a detector. Now they specify a color for each detector setting *and a setting for each detector.* Instead of 8 internal states of the particle, we now have 48 internal states of the particle. The particle state now uniquely determines the color at each detector. Given the (new) particle state, the color at either detector provides no further information about the color at the other detector, because there is no more information to provide. We can give another causal diagram:

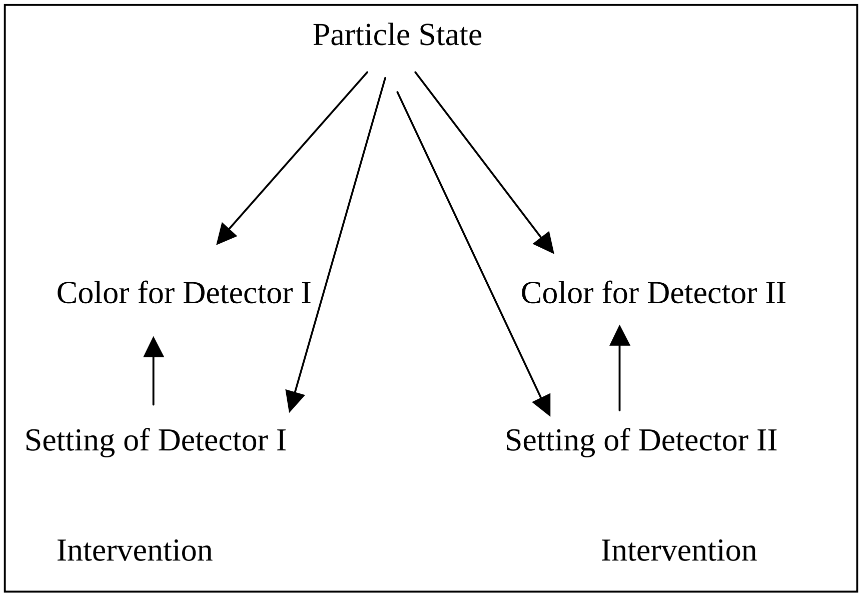


Figure 9

The Markov Assumption is satisfied. (Alternatively, the particle states can influence the interventions, which influence the detector settings.) Why doesn't Mirmin allow this? Because he thinks, quite reasonably, that the particle states do not cause the detector settings. Why not? Because he thinks the human act of setting the detectors (or a machine act of randomly setting the detectors is an *intervention,* a cause that is not influenced by any feature of the system and that fixes the value of the Detector setting while leaving all of the conditional probabilities of other variables unchanged. (Similar reasoning applies to the idea that the detector settings influence the particle state.)

Ok, take out the causal influence of the particle states on the detector settings, but leave the 48 states of the particle and their probabilities just as before:

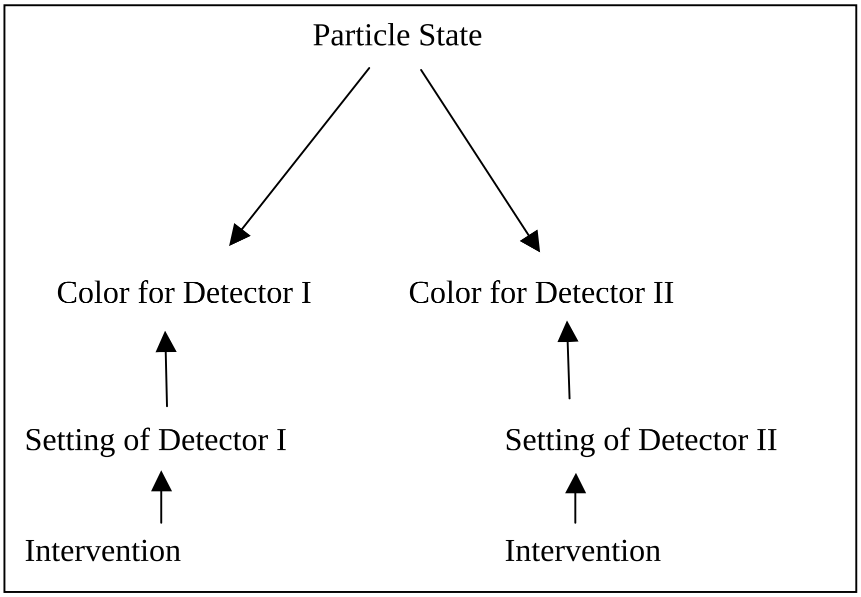


Figure 10

Now we can still account for the correlations in Table 1, and the particle state is still a common cause of the detector colors, condition on which the detector colors are independent—the Markov Assumption is satisfied. Why doesn't Mirmin allow that?

Because the causal diagram in figure 11 and the probabilities assumed for the particle states are jointly inconsistent with the Markov assumption in another way—each detector setting is dependent in probability on the particle state (and vice-versa), but there is no causal pathway or common cause relating the detector setting variables to the particle state. Supposing there is another common cause beside the particle state that also influences the colors won't help things—the same argument goes through, its just more complicated. However, we do things, we do not have a causal explanation of the experiment consistent all the way through with the Markov assumption.

Mirmin—and we—reason about his imaginary experiment using the Markov Assumption and the notion of an intervention, and yet the experiment allows of no causal explanation consistent with the Markov Assumption. The example is a simplification of what goes on in real experiments to test remote correlations predicted by a consequence of the quantum theory, Bell's theorem. In quantum experiments, we pull ourselves *down* by our bootstraps.

Now there is an obvious solution to the problem: the color at one or both of the detectors influences the color at the other detector.

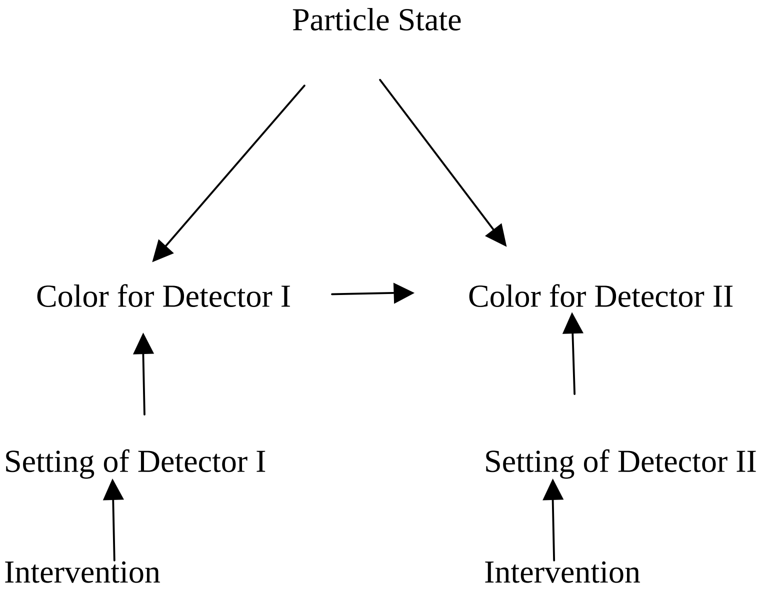


Figure 11

This is a popular solution, and the reason why the problem is often said to be about "locality" or the phenomenon is said to exhibit "non-locality." Often the non-locality solution is implicitly motivated by the idea that the correlations between the colors must have a causal explanation.

Since the detectors can be far enough apart, and the color measurements close enough in time that the theory of relativity prohibits a signal from being sent from one detector to another, the solution has a problem. The problem is this: Suppose before the experiment, the guy at Detector II tells the guy at Detector I how Detector II will be set. Then, if the causal story above is correct, by adjusting the settings of Detector I the first guy can send signals to the second guy, who will figure them out from the color that shows up at Detector II. It works this way. There is in figure 12 a causal pathway from setting of Detector I to the color at Detector II. The pathway must create an association between the two, and associations are all that is needed for communication, for sending a signal. The Faithfulness assumption says a direct causal connection creates an association—and the very point of the non-locality hypothesis is to create such an association between the colors. (Consistently with the Markov Assumtion the association cannot be the effect of a common cause—for reasons we have already reviewed.) The setting of Detector I influences the color at Detector I, so we have a sequence of causal links—and correlations or associations—between Detector 1 and the color at Detector II. Now; a causal linkage of one variable with a second linked with a third need not always create an association between the two variables, even if it is the only pathway connecting the variables (as in this case between Detector I and the color at Detector II). For example, suppose variable A has three values and variable B has three values (say, b1, b2 and b3) and variable C has two values, and the probabilities for two of the values (b1 and b2) of B depend on the value of A, (but the third value, b3, of B does not depend on the value of A) and the probability of values of C depends on whether B has value b3 or one of the values b1, b2, but doesn't depend on which of the values b1 or b2 B has. Then interventions that vary A will not create any association with C. Despite the fact that A influences B, and B influences C, A does not influence C: causation is not transitive. *But* *if B has only two values, the causal relations must be transitive, and A must be associated with C.* That is exactly the situation in the Mermin's thought experiment. Hence relativity can be violated. Having the influence go in both ways doesn't help; the argument still works.

The argument doesn't depend on any philosophical niceties about what "causation" means, and it doesn't depend on any details of the physics. It depends on the assumption that the settings of the Detetctors are interventions, and the hypothesis that the "non-locality" relation creates an influence between the colors. So, if relativity is true and the statistics drawn from the Aspect and similar experiments are sound, causal non-locality is a non-starter.

The upshot is this: real experiments with associations analogous to those of Mermin's thought experiment create associations that have no causal explanation consistent with the Markov Assumption, and the Markov assumption must be applied , implicitly or explicitly, to obtain that conclusion. You can say that there is no causal explanation of the phenomenon, or that there is a causal explanation but it doesn't satisfy the Markov Assumption. I have no trouble with either alternative. It is not a truth of logic that all experimental associations have a causal explanation, and it is not a truth of logic that all causal relations satisfy the Markov Assumption. That's up to Nature. But I do have this problem: *why, then, does the Markov Assumption work with our experiments on middle sized dry and wet goods, with climate, and rats and drugs, and so much else?*

The answer might be quite banal, or quite profound. A banal answer would be that the non-Markovian associations found in the Aspect experiment are so rare in nature, or generally so small, that everything works just fine if we ignore them for experiments and observations in the Large. If I knew a profound answer, I would tell you.

1.2. The Problems of the Aggregate

Another ancient puzzle is the problem of the heap: when does a piling stones one on the other turn the collection into a heap? We have different problems, and they are not about vagueness, but about aggregation.

In macroeconomics, causal relations in an economy are claimed among variables such as price and money supply, GNP and interest rates, interest rates and unemployment rates, and so on. These quantities are aggregated from features of lesser entities, individuals, banks, and such. In ecology, quantitative relations are postulated between such things as predator population size and prey population size, but these quantities are also aggregates. Individual foxes eat individual rabbits, and individual rabbits mate with each other; populations don't eat or mate. Medicine used to think that cholesterol in the arteries causes heart attacks, but it doesn't (so they say now): cholesterol is constituted of two kinds of substances, like jade, and one of the causes heart attacks and the other does not.

Here are the questions: *how do stable statistical regularities at the "macro" level arise from the statistical regularities at the "micro" (where micro = large in the previous section) level? If the real causal relations are at the micro level, how can we predict the effects of interventions which we describe at the macro level?*

There are some interesting positive answers, which I think are not true, some interesting negative answers that are true, and a lot of unsolved problems around these questions. I will start with a positive answer that I think is not true.

1.2.1 Conditional Dependence of the Micro Given the Macro

Consider some randomizing device, a coin flipper say. The coin is made up of a metal lattice of nickel atoms, and each atom has a probability distribution for its location. How does the stable probability of heads arise? One thought is that it works something like this: We know certain macroscopic features of the nickel to be flipped—its parts are bound together spatially, its center of mass is, well, at the center, it is the Earth's gravitational field, and so on. Given those constraints—conditional on them—the particle locations are independent of one another in probability. The probabilitiy that an atom is on the head surface is independent of the probability than some other atom in the nickel is somewhere else. Given the conditional independence of the particles, we can derive a conditional probability for the whole nickel turning up heads or tails when flipped. Similarly, given features of populations of rabbits and foxes, we can suppose the probability individual rabbits are eaten, or not, is independent of the number and locations of other rabbits and foxes, and then we can derive a stable probability relation between rabbit population size and fox population size. I leave aside the details of how this can be done.

This is the idea developed in Michael Strevens' book, *Bigger than Chaos*. I think it is a wonderful book, not for its solution, which I am about to dispute, but for bringing out the problem and bravely attacking it. That's progress.

Suppose we have a bunch of variables, number of foxes per acre say, and number of rabbits per acre, and any macroscopic variable R, say the amount of vegetation and the total population of other vegetarian species in the total region. Strevens strategy is to condition the number of foxes, respectively rabbits, in each acre on some macroscopic property, in this case we will suppose R, and hold that the probability of any number of foxes (respectively rabbits) in an acre is, conditional on the value of R, independent of the number in any other acre. We can work it out this way: the probability of so many rabbits and foxes here and so many there, altogether, still conditional on R, is the product of the individual probabilities for each acre. The logarithm of the product is then the sum of the logarithms of the factors. The sums of independent variables tend toward a nice probability distribution, the chi-square distribution. So we have a joint probability distribution for the total number of foxes and for the total number of rabbits, and we can discover a relation between the marginal distributions (the probability of any number of foxes and the probability of any number of rabbits).

So what's not to like? This: *If the probability of the value of the macroscopic variable, R, whatever it may be, is influenced by the values of the number of foxes and rabbits in each acre, then the fox and rabbit population size in one acre cannot, almost surely, be independent of the fox and rabbit population in other acres.*

Once again, a causal diagram:

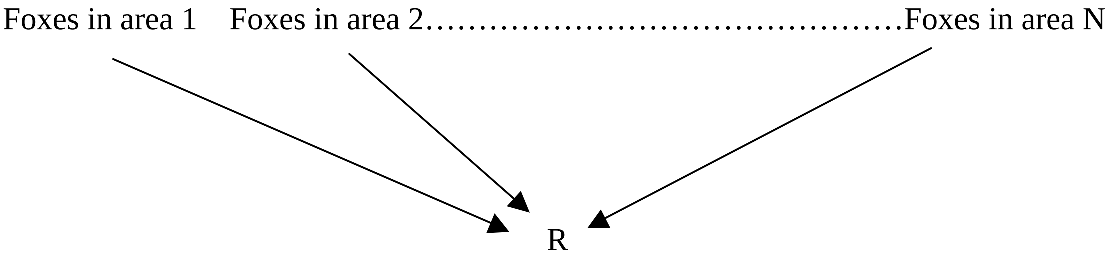


Figure 12

For all joint probability distributions on the variables, the fox variables will be dependent conditional on R, even if they are independent unconditionally. Faithfulness requires it, and while it is easy to manufacture particular unfaithful probability distributions, they are unstable—all the distributions close to them satisfying the Markov Assumption are faithful. So Strevens' idea might work in particular cases, but it won't work in general, or, I should think, often.

1.2.2 Conditional Dependence of the Macro Given the Micro

Conditional probability creates problems in the other directions as well. Sometimes we want to learn about micro properties of individuals from aggregations of their properties. When studying how genes regulate one another, for example, DNA in the nucleus of a cell is copied or transcribed into messenger RNA, which is then spliced into pieces by enzymes; some of the pieces migrate to a protein complex which holds as if in a frame, and there, using the RNA pieces as a template, amino acids are stitched together to form proteins. Some of the proteins then regulate how other genes (or the one from which they originated even) are transcribed. Genes regulate other genes through a cascade of such effects, often started or modulated by other factors—small molecules, for example. Understanding how this all works is a big project in cell biology.

One way of studying gene regulation is to measure the mRNA concentration in a bunch of cells from a tissue under varying conditions—one can knock out or inhibit the expression of a particular gene for example, and see how the mRNA products of other genes change. Or one can study the correlations among mRNA concentrations for various genes and from the correlations try to figure out which genes are regulating which other genes directly or by regulating genes that in turn regulate other genes. Unfortunately, it doesn’t work.

Suppose in each cell in a sample, the expression of Gene 1 influences expression of Gene 2 which influences expression of Gene 3, and expression of Gene 3 is also influenced by expression of Gene 4. We cannot measure the concentrations of messenger RNA in each cell, only the aggregate concentrations in a great many cells. Our aggregated measurements give us only two kinds of clues to the regulatory structure in the cells: the correlations of each pair of the measured concentrations, and the correlations of any pair conditional on the measured concentrations of the mRNA from other genes. There is nothing wrong with the correlations, although we might need a very large correlation or a very large number of repetitions of the experiment to separate signal from noise. But the correlations don’t tell us much. If the aggregated mRNA concentrations of genes A and B are correlated, that might be because A regulates B directly, or B regulates A directly, or one of them regulates the other through some intermediate gene, or some other gene regulates both A and B.

To get more information, we have to consult the conditional associations. If gene A expression is correlated with gene C expression in our aggregated measures, but these are not independent conditional on expression of gene B, then by the Markov and Faithfulness Assumptions, we should conclude that B is not an intermediate between A and C. We would be wrong. If we conducted an experiment in which we perturbed gene

A, say, and the expression of B and C varied in response, but the variation in C was not independent of the variation in a conditional on B, we might apply the Markov Assumption to conclude that A causes B and C, and B is not an intermediate. We would be wrong.

The error is not in the Markov or Faithfulness Assumptions, but in applying them to aggregated variables that are not causes of one another, but emanations of the underlying causal relations. Suppose, again, the regulatory structure at the cellular level is:

A → B → C ← D

Suppose C is a nonlinear function of B and D. Then the Markov Assumption says that A and C are independent conditional on B, and at the level of the individual cell, they are. But our measurements are of sum (or concentration) of mRNA products for A, B, C, and D for a great many cells. And the sum of A values is not independent of the sum of C values conditional on the sum of B values. It turns out that gene regulation is like that—non-linear functions all over the place. We cannot, except in special cases of linear systems or systems with Normally distributed variables, extract the causal structure of the microcosm from observations of conditional associations in the macrocosm.

So what? So there is something important we can’t learn in a particular way. What is the philosophical issue? This, I think: Other than observing correlations and time order, we can learn macroscopic causal relations from observation—as in astronomy and geology and social science and elsewhere--only from conditional independence relations among macroscopic features. There is a general theorem, due to Tianjaio Chu, that essentially says that conditional independence relations among microscopic variables are preserved in their macroscopic sums, only if the dependencies are linear or Normal. The internal properties of middle sized dry and wet goods, and their apparent causal relations with one another are produced by their microscopic components and their causal relations with one another. The relevant relations of component properties are not generally linear. *So why* *are there any conditional independence relations at all among macroscopic features of the world?*

1.2.3 Knowing the Big and Manipulating the Small[1]

When an experiment is done, something is manipulated, a drug known to reduce cholesterol is given for example, and something happens in response. Perhaps heart attacks decline. But in response to what? The drug, of course, but what about the drug, what does the drug do? Reduces cholesterol, of course. Maybe not. Imagine the following scenario:

Experimenter 1 gives drug CholA to a random sample of patients and a placebo;
Experimenter 2 gives drug CholdB to different groups the same way;
Experimenter 3 gives drug CholC to different group the same way
Experimenter 4 gives drug CholD to different group the same way

All four drugs are known to reduce cholesterol. Experiment 1 finds a small association between drug treatment and reduced heart attacks; Experiment 2 does the same; Experiment 3 finds no association, and Experiment 4, with a larger sample size than the others, finds a reverse association—the drug increased heart.attacks. All of the experiments show the drugs reduced cholesterol. A meta-analysis, which effectively treats all of the samples as though they were a single sample, finds no association.

What can have gone wrong? The first two of the drugs reduced cholesterol by reducing low density and high density cholesterol; the third drug only reduced low density cholesterol; ; and the fourth drug only reduced high density cholesterol. Only low density cholesterol causes heart attacks. Each experimenter thought the treatment was a manipulation of “cholesterol” in arteries, but cholesterol is the sum of two kinds of cholesterol, and some the treatments manipulated the kinds differently, with different results.

[1] This section draws on work by my colleagues, Peter Spirtes and Richard Scheines.

This is not just a problem about discovering causal relations and their mechanisms, although it is that. It is also a problem about understanding how to describe and to represent causal relations, and how to predict the results of interventions or experiments when we think we know the causal relations. The issues arise in epidemiology, in economics, indeed anywhere that our causal explanations are in terms of variables that are aggregated from other variables that are the features that, wittingly or unwittingly, we singularly intervene on. I will look at our example in more detail.

Through correlations, researchers discover, they think, that high cholesterol levels cause heart disease. They recommend lower cholesterol diets to prevent heart disease. But, unknown to them, there are two sorts of cholesterol: one sort of cholesterol causes heart disease, the other does not. Low cholesterol diets differ, and they differ in the proportions of the two kinds of cholesterol. Consequently, experiments with low cholesterol regimens differ considerably in their outcomes.

In such a case the variable identified as causal—total cholesterol—is singularly a deterministic function of two underlying factors, one of which is singularly causal, the other not. The interventions (diets) are singularly interventions on the underlying factors, but in different proportions. How are such causal relations to be represented, when do the Markov and/or Faithfulness assumptions hold and when do they not, and how should one conduct search when the systems under study may, for all one knows, have this sort of hidden structure? These issues seem important to understanding possible reasons for disagreements between observational and experimental studies, non-repeatability of experimental studies (and not only in medicine—psychology present many examples), and in understanding the value and limitations of meta-analysis.

Consider what kinds of dependency structures can emerge in a few hypothetical examples. In the hypothetical Example 1, shown in Figure ??, the concentration of total cholesterol is defined in terms of the concentrations of high density lipids and low density lipids. This is indicated in the figure by the bold faced arrows from *HDL* and *LDL* to *TC*.

The other arrows indicate causal relationships. Suppose that high levels of *HDL* tend to prevent *HD*, while high levels of *LDL* tend to cause *HD*. We have the following parameters for Example 1.

HDL = Low, *LDL* = Low → *TC* = Low

HDL = Low, *LDL* = High → *TC* = Medium

HDL = High, *LDL* = Low → *TC* = Medium

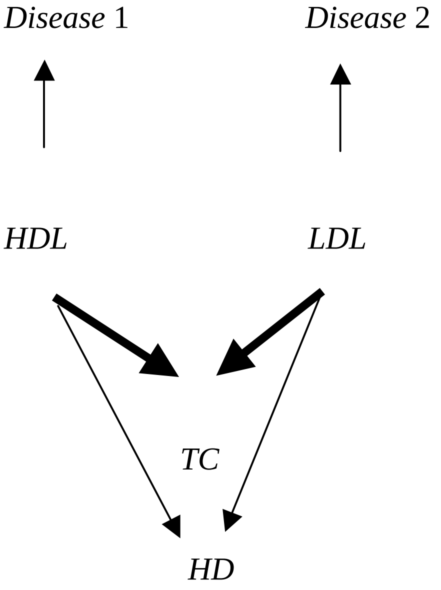
HDL = High, *LDL* = High → *TC* = High

P(*HDL* = High) = .2
P(*LDL* = High) = .4
P(*Disease* 1 = Present|*HDL* = Low) = .2
P(*Disease* 1 = Present|*HDL* = High) = .9
P(*Disease* 2 = Present|*LDL* = Low) = .3
P(*Disease* 2 = Present|*LDL* = High) = .8

P(*HD* = Present|*HDL* = Low, *LDL* = Low) = .4 = P(*HD* = Present|*TC* = Low)
P(*HD* = Present|*HDL* = High, *LDL* = Low) = .1
P(*HD* = Present|*HDL* = Low, *LDL* = High) = .8
P(*HD* = Present|*HDL* = High, *LDL* = High) = .3 = P(*HD* = Present|*TC* = High)

Manipulation of *TC* is really a manipulation of *HDL* and *LDL*. However, even after an exact level of *TC* is specified as the target of a manipulation, there are different possible manipulations of *HDL* and *LDL* compatible with that target. For example, if a manipulation sets *TC* to Medium, then this could be produced by manipulating *HDL* to Low and *LDL* to High, or by manipulating *HDL* to High and *LDL* to Low. Thus, even after the manipulation of *TC* is completely specified (e.g. to Medium), the effect of the manipulation on *HD* is indeterminate (i.e. if the manipulation is *HDL* to High and *LDL* to Low, then after the manipulation P(*HD*) is .1, but if the manipulation is *HDL* to Low and *HDL* to High, then after the manipulation P(*HD*) is .8). Hence a manipulation of *TC* to Medium might either lower the probability of *HD* (compared to the population rate), or it might raise the probability of *HD*. It is quite plausible that in many instances, someone performing a manipulation upon *TC* would not know about the existence of the underlying variables *HDL* and *LDL*, and would not know that the manipulation they performed was ambiguous with respect to underlying variables. For example, manipulation of *TC* could be produced by the administration of several different drugs which affect *HDL* and *LDL* in different ways, and produce different effects on *HD*.

What is the correct answer to the question “What is the effect of manipulating *TC* to Medium on *HD*?” The most informative answer that could be given is to give the entire range of effects of manipulating *TC* to Medium (i.e. either P(*HD*) = .8 or P(*HD* = .1)). Another possible answer is to simply output “Can’t tell” because the answer is indeterminate from the information given. A third, but misleading, answer would be to output one of the many possible answers (e.g. P(*HD*) = .1).This answer is misleading as long as it contains no indication that this is merely one of a set of possible different answers, and an singular manipulation of *TC* to Medium might lead to a completely different result. Note that it is the third, misleading, kind of answer that would be produced by performing a randomized clinical trial on *TC*; there would be nothing in the trial to indicate that the results of the trial depended crucially upon details of how the manipulation was done.



HDL: High Density Lipids
LDL: Low Density Lipids
TC: Total Cholestorol
HD: Heart Disease

Figure 13

Suppose now that *Disease* 1, *Disease* 2, *TC*, and *HD* are the measured variables, and we assume the Markov and Faithfulness Conditions (extended to graphs with definitional links), but allow that there may be hidden common causes. What (pointwise consistent) inferences can be drawn from large samples of the distribution described in Example 1? We will contrast two cases: the case where it is assumed that all manipulations are unambiguous manipulations of underlying variables to the case where the possibility that a manipulation may be ambiguous is allowed. The general effect of weakening the assumption of no ambiguous manipulations is to introduce more "Can't tell" entries.

Table 3

Manipulate:	Effect on:	Assume manipulation unambiguous	Manipulation may be ambiguous
Disease 1	*Disease* 2	None	None
Disease 1	*HD*	Can't tell	Can't tell
Disease 1	*TC*	Can't tell	Can't tell
Disease 2	*Disease* 1	None	None
Disease 2	*HD*	Can't tell	Can't tell
Disease 2	*TC*	Can't tell	Can't tell
TC	*Disease* 1	None	Can't tell
TC	*Disease* 2	None	Can't tell
TC	*HD*	Can't tell	Can't tell
HD	*Disease* 1	None	Can't tell
HD	*Disease* 2	None	Can't tell
HD	*TC*	Can't tell	Can't tell

We will now consider what happens when the example is changed slightly. In Example 2, suppose that the effect of *HDL* and *LDL* on *HD* singularly is completely determined by *TC*.

Example 2 is the same as the Example 1, except that we have changed the distribution of *HD* in the following way:

P(*HD* = Present|*HDL* = Low, *LDL* = Low) = .1 = P(*HD* = Present|*TC* = Low)
P(*HD* = Present|*HDL* = High, *LDL* = Low) = P(*HD* = Present|*HDL* = Low, *LDL* = High)
= .3 = P(*HD* = Present|*TC* = Medium)
P(*HD* = Present|*HDL* = High, *LDL* = High) = .8 = P(*HD* = Present|*TC* = High)

In this case, while manipulating *TC* to Medium represents several different manipulations of the underlying variables *HDL* and *LDL*, each of the different manipulations of *HDL* and *LDL* compatible with manipulating *TC* to Medium produces the same effect on *HD* (i.e. P(*HD*) after manipulation is equal to P(*HD* = Present|*HDL* = High, *LDL* = Low) = P(*HD* = Present|*HDL* = Low, *LDL* = High) prior to manipulation, which is .3). In this case we say that the effect of manipulating *TC* on *HDL* is determinate. (Note that the effect of manipulating *TC* on *Disease* 1 is not determinate, because it depends upon how the manipulation of *TC* is done. So manipulating a variable may have determinate effects on some variables, but not on others.)

The Faithfulness assumption that we have been making singularly entails that the effect of *TC* on *HD* is not determinate. (This is because if the effect of manipulating *TC* on *HD* is determinate, then *LDL* and *HDL* are independent of *HD* conditional on *TC*, which is not entailed by the structure of the causal graph, but instead holds only for certain values of the parameters, i.e. those values for which P(*HD* = Present|*HD* = Low, *LDL* = High) = P(*HD* = Present|*HDL* = High, *LDL* = Low).) Hence, in these cases we make a modified version of the Faithfulness Assumption. What (pointwise consistent) inferences can be drawn from large samples of the distribution described in Example 2? Because there are conditional independence relations that hold in Example 2 that do not hold in Example 1, more pointwise consistent estimates of manipulated quantities can be made under the assumption that manipulations may be ambiguous, than could be made in the previous example.

Table 4

Manipulate:	Effect on:	Assume manipulation unambiguous: Example 2	Manipulation may be ambiguous: Example 2
Disease 1	*Disease* 2	None	None
Disease 1	*HD*	Can't tell	Can't tell
Disease 1	*TC*	Can't tell	Can't tell
Disease 2	*Disease* 1	None	None
Disease 2	*HD*	Can't tell	Can't tell
Disease 2	*TC*	Can't tell	Can't tell
TC	*Disease* 1	None	Can't tell
TC	*Disease* 2	None	Can't tell
TC	*HD*	= P(*HD*|*TC*)	= P(*HD*|*TC*)
HD	*Disease* 1	None	None
HD	*Disease* 2	None	None
HD	*TC*	None	None

Example 1 and Example 2 are two simple cases in which causal conclusions can be reliably made. Indeed, for those examples, the algorithms that we have already developed and that are reliable under the assumption that there are no ambiguous manipulations, still give correct output, as long as the output is suitably reinterpreted according to some simple rules that only slightly weaken the conclusions that can be drawn. However, there are other examples in which this is not the case. For example, if *Disease* 1 and *Disease* 2 are not independent, but are independent conditional on a third measured variable *X* then no simple reinterpretation of the output of the algorithm gives answers which are both informative about cases in which *TC* does determinately cause *HD*, and reliable. In all such examples that we have examined so far, however, the data itself contains information which indicates that the current algorithm cannot be applied reliably; hence for these examples the algorithm could simply be modified to check the data for this condition, and output “can’t tell.” But we do not have general conditions under which the data would indicate that the algorithm cannot be reliably applied (unless the assumption of no ambiguous manipulations is made.) This raises the questions: Are there feasible general algorithms that are both correct and informative even when the assumption of no ambiguous manipulations is not made? If so, what is the algorithm? What is its computational complexity as a function of the number of variables? What is its reliability on various sample sizes?

References

P. Spirtes, C. Glymour and R. Scheines, Causation, Prediction and Search, Springer Lecture Notes in Statistics, 1993; 2nd edition, MIT Press, 2000.

N. David Mermin, *Boojums All the Way Through*, Cambridge University Press, 1990.

T. Maudlin, *Quantum Non-Locality and Relativity*, Blackwell, 1994.

M. Strevens, Bigger than Chaos, Harvard University Press, 2002.

T. Chu, C. Glymour, R. Scheines and P. Spirtes, A Statistical Problem for Inference to Regulatory Structure from Measurements of Gene Expression with Microarrays, *Bioinformatics*, 19, 1 – 6, 2003.

P. Spirtes, P, and R. Scheines,. Causal Inference of Ambiguous Manipulations (to appear). *Proceedings of the 2002 Philosophy of Science Association Meetings.*