DRAFT of Adam Morton, *Acting to Know: the epistemology of experiment*.

CONTENTS

preface

In this book I develop and defend a novel account of evidence. Evidence supports a hypothesis, on this account, by putting on the path to knowledge. It applies best to evidence derived from experiment, and it is indeed intended to express what is special about this source of evidence. It applies ideas from contemporary epistemology to a complex of problems that are usually the concern of statistics and the philosophy of science, centering on the support that scientific practices give to hypotheses. There are three main themes, which can only be defended, and perhaps only stated, by combining ideas from several sources in un-traditional ways. They are:

– a general description of nearness to knowledge that applies in a number of areas well beyond that of known claims, conclusions, theories, propositions and the like.

– the use of this to say what evidence is, and how it supports claims.

- a discussion of statistical tests that draws on ideas about possibility and combines them with ideas about probability

- a concept of knowledge, or something knowledge-like, that comes in degrees of strength, at the weak end so weak that satisfying it would never justify a normal ascription of knowledge

As I shall develop these ideas they are in conflict with several dominant views in epistemology and the philosophy of science. Chapter 0 gives a foretaste of this and chapter 8 puts the pieces together into a single coherent position. Each of the chapters in between discusses a single topic in isolation in a form that I hope will stand on its own, avoiding a house of cards construction where a single failure can bring down the whole

business and allowing readers who are unconvinced by some claims to be persuaded by others.

Fasten a pair of calipers tightly around an object. You can then read how wide and often how solid it is. If you come to think that your measurement was wrong you know how to go about repeating it, if need be with an improved tool. This gives a model of a certain kind of information-gathering. We interact causally with things, in a way that allows us to change, correct, and expand our information. The information doesn't just pile up, as a storehouse of items that can be true or false of their objects. Rather, the causal process that provides information also gives us ways of correcting, improving, and extending it. And the capacity to do this is in some respects more important than the accumulation.

While thinking and writing about experiment and evidence I did a series of interviews with experimenters at UBC. These were friendly and interesting occasions and I learned a lot from them. I am grateful for many people's help. They include Holly Andersen, Marcel Bally, Prasanta Bandyopadhyay, Jess Brewer, Matthew Cobb, Allan Franklin, Clark Glymour, Madelyn Glymour, Jason Grossman, Francesco Guala, Kiley Hamlin, Parisa Mehrkhodavandi, Sonia Memetea, Slobodan Perovic, John Petkau, Margaret Schabas, Toni Schmader, Trish Schulte, George E Smith, Dan Steel, Douw Steyn, Mike Whitlock, and Karen Zwier. Alirio Rosales has been a constant source of advice and comment. Susanna Braund wanted me to write a rather different book, so I hope she is not too disappointed in the one I was able to write.

The link between causation and experiment plays an important role in the work of two

philosophers who I take to be arguing for positions that resemble mine, but which I shall only discuss in passing. Ian Hacking's *Representing and Intervening* has inspired sociologists and historians of science to pay attention to experimentation as a basic scientific activity with its own culture. This book convinced me when I first read and indeed reviewed it that issues about realism need to be approached in terms of scientific activity as much as scientific doctrine. But its emphasis is not on evidence so there are not many direct links with my themes in this book. Similarly with James Woodward's work. Woodward says very little about the connection between experimenting and evidence, and my approach to causation is motivated by that connection. In order to discuss him I would have to supply connections that might well not be his or which he might even oppose. So I say nothing substantive about either Hacking or Woodward although they have both influenced my thinking about experiment.

chapter 0: two evidential strategies

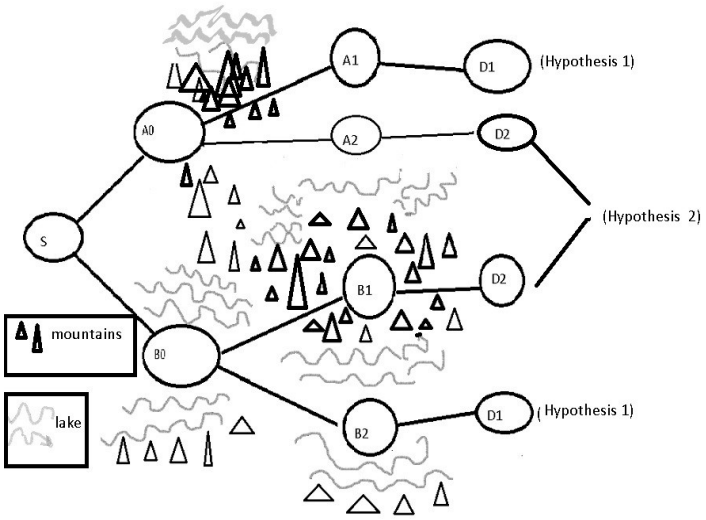
This mini-chapter introduces a distinction about evidence that will underlie what follows. The subsequent chapters refine it and defend its usefulness. But they aim at a number of component issues, which are not put together until the final chapter 7. So to give some perspective here is a glance ahead. The chapter is constructed around an example illustrating the distinction. It contrasts two general strategies for supporting a hypothesis. One strategy understands the relation between a hypothesis and evidence for it as similar to the relation between states of mind and the environment when one has knowledge.[1] The other strategy understands it in terms of norms or standard procedures of rational belief-formation. Various forms of this second strategy are usually taken for granted in discussions of evidence, but I am defending the first one and contrasting the two. Its advantages are greatest when evidence is collected by experiment

the train case

Two passengers, Sophia and Norm, are on a train that has gone through a series of branching junctions. They fall asleep and wake feeling that they have slept for a long time and with doubts about whether they are on the right train. Their aim had been to get to one destination (D1) but they fear that they are by mistake on a train heading for another (D2). So they look out the windows, consider the landscape, and compare it to

1 Vaguely "environment" to avoid building propositions or facts into the ontology. Complexes of possible worlds would probably do but I want a minimum of orthogonal issues in this connection. Propositional attitudes such as belief are hard enough to understand in relational terms, but factive attitudes such as knowledge, which require a really existing "object" are even more puzzling.

what they know about the route. They see near mountains and further away a large lake. This could fit a number of places along the way (see the diagram).



Assuming that they have slept for a good while the superficially best fit is with B1, which would mean that S and N are headed for D2. But in fact, what they thought was a long sleep was just a nap and they are leaving A0, so headed for D1. Here are two ways of deciding where they think they are, and for each a corresponding way of evaluating the resulting belief.

(a) realist, Sophia's strategy: look out the train window, use the observation to decide between two previously chosen alternatives. Count the result a success if it gives a true belief, and in circumstances where you would have chosen its alternative that alternative would have been true instead.

(b) norm-based, Norm's strategy: look out of the train window and use what you see in reasoning leading to whichever hypothesis it makes more probable given your background information, notably about probabilities. Count the result a success if it gives a true belief in this and other probable situations.

The relevant difference between (a) and (b) is that on (a) possibilities that are nearer to actuality (in this case, require fewer branchings away from the actual history) are central while on (b) one prioritizes the more probable possibilities, given what else one believes.

There are obviously other ways to form beliefs in situations like this and other ways to evaluate the results. My interest in these is as contrasting two evaluations of the force of evidence. The realist method will in this case deliver hypothesis 1, and will count this as a success. For in fact we have just passed A0 and are thus on the way to D1, as hypothesis 1 asserts. Moreover in the nearest alternative situation where the method would have given the alternative conclusion, that we were headed for D2, hypothesis 2 would have been true, since we would be around B1.

The norm-based method will fail in this case; Norm's conclusion is false and Sophia's is true We think it most probable that we have slept for a while, and the most likely mountains-then-lake scenery given this is along the central two routes, suggesting the false hypothesis 2.

The reason for the failure of the norm-based method is its use of wrong probabilities. (We think that we have probably slept for a long time.) The reasons for the success of

the realist method are that it centres on the actual situation, whether or not the person can describe its relevant features, and compares the two alternatives in terms of their treatment in the nearest situations where they would be chosen, in this case those that are least distant and involve fewest branchings from the starting point. These are possible situations that require changing causal features of the actual situation least, so that the result is evaluated in terms of its treatment of objectively similar situations.

knowledge and evidence

A main aim of this book is to make a case for ways of evaluating evidence generally like Sophia's objective method above. A single example does not show much, as the method might have been specially fitted to fit the example. A general justification is needed. Experiment is closely related to it, and gives some of the clearest and most convincing cases. I described it as if our travellers were accepting one hypothesis or the other on the basis of what they saw from the train window. The criterion for knowledge was along the lines of what are standardly called "safety" considerations.[2] But we might more generally be concerned with which hypothesis the evidence supported best, even if neither was supported well enough to be a candidate for knowledge. In the example the similarity between full knowledge and the situation of the hypothesis given data is not hard to describe. Substitute "prefer" for "accept" throughout, "would have been nearer to truth (in the nearest situation)" for "would have been true" in the realist method and "more often true" (in similar situations) in the norm-based method: the result is not a criterion of knowledge but one of evidence. This is what I shall call *K-evidence* and slowly

2 Contrasted with "sensitivity" considerations. The contrast between the two is however minimal when we are choosing between two incompatible hypotheses. See Chapters 6 and 7.

characterize, contrasting it with the norm-based *R-evidence.*[3] Then the pieces are assembled in the conclusion chapter.

Two contrasts between the strategies are particularly important (for my purposes, at any rate). The first is that K-evidence is independent of the agents' beliefs, prior knowledge, and the like, and of what is rational for or according to agents. Indeed, one can have this kind of evidence without knowing that one has it, and without knowing how it compares to the evidence one has or would have for another hypothesis or given different data. This is so for generally the same reasons that one can know or fail to know without realizing that what one has is or is not knowledge, and that one will usually not assess the extent of one's knowledge or ignorance accurately. The second contrast is between possibility and probability. As I am construing K-evidence it makes essential use of what can or might happen or be the case, understanding this as a generally speaking causal notion — what can occur given that the world works the way it does — and moreover of degrees of possibility as expressed in terms of nearer and more remote situations or possible worlds. The versions of the norm view that will concern me most, in contrast, make essential use of the concept of probability. I shall argue that this has to be understood as itself a generally causal concept.

A connection between the two: K-evidence indicates the range of similar circumstances a hypothesis holds, and in which action based on the hypothesis will usually succeed. N-evidence indicates factors normally but not necessarily correlated with this range and

3 Part of the wider importance of the norms/knowledge contrast is the fact that statistical ways of collecting and evaluating evidence seem arbitrary and alien to many. Why defer to *these*? Objective standards of evidence allow the beginnings of answers. But issues about the public perception of science are not central to this project.

typically useful as inputs to standard patterns of reasoning, whose reliability can vary according to features of the situation of which the person may not be aware.

Each way of understanding evidence has its advantages. My job is to sing those of the underappreciated objective approach. A disadvantage of the norm-based approach is the potential indefiniteness of its targets, and the associated endless list of norms that would have to be considered. Consider the variety of factors that one could rationally take account of. The most immediate is available data that is easily understood and already rich and varied. But beyond this there is evidence of further data not in one's possession. This comes in a number of forms, which are relevant to belief in a number of ways.[4] Further beyond there are reasons to think that more information can be got from the available data than one has managed to extract. And often there is reason to take account of one's own likely failures to interpret data.[5] In quite a different direction there are principles of not wasting time and thinking power and not over-scrutinizing evidence (understanding and following which can waste a lot of time and thinking power). Which of these constitute norms of reasonable evidence? In what ways are they similar? It seems that even beginning to think in this way one is being led into a labyrinth.

experiment

4 Christensen (2010), Tal and Comesaña (2017)

5 In Morton (2012) I develop an attitude to issues of human fallibility and finiteness, based on the concept of an epistemic virtue. These topics will feature very little in this book. A simple connection between epistemic virtues and K-evidence is that sensitivity to K-evidence of some kind, including evidence one is not conscious of having, is an epistemic virtue in that it is disposes one towards corresponding true beliefs.

Both kinds of evidence are available to both Sophia and Norm, though each will use just one to shape their beliefs. Norm may not be aware of the force of the available objective evidence. Indeed Sophia may be ignorant or mistaken about it also. Some degree of inaccuracy is almost inevitable, and to that extent her grasp of the force of her evidence and the reaction to it is likely to be somewhat rough.

When the source of objective evidence is experiment, though, this problem is much diminished. Experiments are deliberate and carefully controlled processes, designed to give results of particular kinds for particular reasons. So when you run an experiment you know what evidence has been produced. The lake or the mountain may be too far away for Sophia to see them, so she may use binoculars; it may be dark, so she may shine a light to see if there is a reflective twinkle from the water. Then she will combine the advantages of objective and norm based evidence.

Experiment has other advantages also. The planning and control will make it easier for two or more people to cooperate in producing and assessing the evidence. It will thus allow them to combine their practical and thinking powers. Indeed an experiment often requires the efforts of several or many people. Norm may help Sophia construct and operate her experiment, perhaps because she has designed it she is not good at operating apparatus. Then he is likely to appreciate the reasons supporting her conclusion.

The result is that experiment often leads to knowledge. Of course there is non-shared knowledge also. But experiments often lead to *better* knowledge, knowledge that has its

defining features to a greater degree. The evidence it uses will have more of these advantages than evidence that leads to more marginal knowledge.

modality

As the train navigation example suggests, objective accounts rely on ideas about what can or would occur, particularly what beliefs someone would have in different circumstances and which of these would be true. These are to be understood as real objective facts about a person and her environment, which like all such facts can be very different from what we think they are. The most important modal facts concern causation, and I shall refer to the whole category of concepts as causal. While some philosophers resist the idea that facts about causation and what might occur under various circumstances are independent of our opinions, it would be very bad news for human decision-making if they were not so independent. We often plan in terms of them: if I do this the following will occur; an action of this type would cause a result of this other kind. We obviously need our opinions about these to correspond to what will actually occur, and if they fail too often the cruel world will take its penalty. So there ought to be a presumption in favour of objectivity. The use of causal ideas in decision-making also suggests that they are important in belief formation, since a primary function of our beliefs is to guide our actions.

Causal and modal concepts will play important roles throughout. My way of organizing them will be standard and unoriginal: the possible worlds orthodoxy of propositions, sentences, beliefs and the like, true or false (or holding) in possible worlds (or situations,

or possibilities — simply stylistic variations here). This will often be represented using Lewis's spheres of proximity or in some other way. And I shall assume that there is a relation of nearness between worlds, in particular to the actual world. (The inverse of nearness is remoteness.) The immediate application of nearness is in the standard definition of a counterfactual, or better subjunctive, conditional, which I shall stick to unless otherwise signalled, as true in a world w when the consequent holds in the nearest world or worlds to w where the antecedent holds. Then we can define "the nearest world where p is true is nearer than the nearest world where q is true", as *((pvq) & ~(p & q)) → p* , where → is the counterfactual (if exactly one of them is true it is p). I take it that an intuitive concept of the nearness of possibilities is implicit in our everyday use of the counterfactual, and in idioms such as "if, and it is a big if,…" and "just possibly" or "it is remotely possible that…".[6]

And, possibly most contentious but not defended here, I shall assume that the truth values of counterfactuals are matters of objective fact, at any rate as much as most of our truth-value-receiving claims. "If the incision had been a millimetre to either side she would have died" says that in situations where there is a tiny change in the actual history, such as a tremor in the surgeon's hands, making the incision just different enough, she does not survive. She really would not have; it is a medical fact that might be the basis for legal action and might be explained by some true medical theory. It does not mean that she would have died in more "remote" situations like that where the

6 Some idioms for evoking degrees of possibility are also associated with probability. When discussing what would happen if humans colonized Mars we say they would have long term food production problems, partly because their being wiped out by indigenous intelligence Martians is *less likely* than their confronting a hostile planet with at most primitive life. A common framework for possibility and probability is highly desirable, but I am not offering it.

surgeon hiccups but a force field from outside the operating theatre manoeuvres her organs out of the way.[7]

7 I require only that nearness/remoteness be a partial ordering.

chapter 1: sensitive evidence

kinds of evidence

Some evidence is better than others. But there are many forms this can take. Better evidence can suggest how to get further evidence; it can get us further along from conjecture to acceptance; and its objective connections with the hypothesis it supports can be stronger. The focus of this chapter is to the second and third of these. But they are all linked in obvious and subtle ways. The hope in accepting a belief, hypothesis, or conjecture is to gain knowledge of its subject matter, that is at a minimum to have an account of it that is true and held because of influence of the facts that make it true. (In the past fifty years epistemologists have laboured make precise versions of this rough idea. Those not in the tribe will often think this is just a symptom of obsessiveness; a secondary aim is to show the importance of the project for real inquiry.) Of course, most evidence leads us well short of knowledge. Weak evidence is still evidence. But the connection with knowledge consists in more than a general aim to attain it. Better evidence often has a knowledge-like relation to a hypothesis, as chapter 0 hinted and this chapter aims to explain. This relation is brewed from the same ingredients as knowledge, but differs in that it can be much weaker; it is a connection rather than an attribute, sometimes connecting evidence to conjectures that are still conjectural. I shall say that evidence that has an appreciable degree of this to-be-explained knowledge-like quality with respect to a hypothesis (theory, belief, idea, …) is *sensitive* evidence for it.[8]

8 Sensitive evidence is similar to what philosophers of science call robust evidence (Achinstein 2001, Staley 2004). I am using a different term partly to emphasize the affinity to an overarching concept of knowledge, and partly because the discussion in the philosophy of science is not usually explicit on a crucial aspect for my purposes, that it is a matter of what actually is the case rather than what a scientist, or would-be knower

Evidence derived from well conducted scientific experiments provides core cases of sensitive evidence, although there are instances in everyday life. And in fact the continuity between everyday and scientific uses of gathering such evidence is part of what gives it intuitive force in science and part of what gives anti-sceptical arguments their plausibility. (It also motivates a reason for a kind of scepticism about many everyday beliefs, as I explain in the book's conclusion.)

Sensitive evidence comes in degrees, with respect to both support and progress towards knowledge. Some experiments deliver its advantages more than others. So do some everyday epistemic procedures. How sensitive a particular item of evidence happens to be is, however, often a factual matter that is not easily known. I shall give some examples of varying degrees of sensitivity, to summon the intuitions that I then try to capture.

insensitive evidence

I shall begin with examples of the kind of evidence that is *not* very sensitive.

I am investigating a favourite author, Julia Scriptor. At first I know nothing but that she wrote the book I admire, but then I learn that she was born in 1910. (Suppose that I learn nothing else, and suppose that I learn this simply by chancing across her birth certificate or from a note on the copyright page of a book.) On learning this I think that it is probable that she is no longer alive, and in fact I take this for granted in my further

in general, believes.

research. The statistical probability that she is not with us is very high. So there is an obvious way that the birth information is evidence for the belief that she is dead.

There is something rather indefinite about the evidence, though, in spite of its power to bring a conjecture near to belief. Suppose that in fact Ms Scriptor is still drawing a pension. This will not change the information that she was born in 1910, or the process that brought this information to my attention, so that if I had checked up on it I would not have noticed anything different. Similarly if she has died. The evidence is not sensitive with respect to the conclusion I base on it, in both the crude sense that it would be the same if she has not died, and the more detailed sense that if she has died then the evidence would be the same whether she died a little earlier or later or if the cause of her death were different.

Another everyday example, circumstantial evidence. The accused had a grudge against the victim and stood to gain from his death. Moreover, she has a PhD in biochemistry, and the victim was poisoned. So it is natural to suspect her, and the victim's family are not unreasonable in taking her to be responsible. But again there is something unsatisfying about the evidence, which her defence lawyer would no doubt draw attention to, by saying that it is all "circumstantial". No fingerprints, no witnesses, no DNA, no precise verifiable timeline of her movements. Again, we can get a hold on what this circumstantiality consists in by noting that her crucial actions with respect to the victim (poisoning or not) have no effect in shaping the evidence: it is the way it is whatever she has done. The evidence raises the probability that she is guilty, but is not sensitive to the facts.

Another way of putting it. The process from fact to evidence is not at all like a measuring or imaging device or an experiment. It does not take a situation and transform it into a form that gives different information depending on the nature of the situation. This is a rough suggestive way of putting it. We need a characterization of experiment to know whether the metaphor has any depth.

Now an example from science. The evidence that the extinction of the dinosaurs was caused by the impact of an asteroid consists in the signs of an asteroid impact in Yucatán, and the layer of iridium in many places at about the right stratum to date with the beginning of the dinosaurs' decline as measured in the standard ways in paleontology. No one takes this to be conclusive but it is generally thought to be fairly strong evidence. It is an inference to the best explanation: this hypothesis accounts well for the data. But strong though it is, it is not very sensitive. The path from dinosaur days to now preserves signs of impact and fossils, but not results of asteroid-induced death. Not even signs of starvation or cold. Such evidence is conceivable, but we do not have it. So it is not at all like a chronoscopic view of the relevant events, but a transmission of other bits of information, which we can put together and add up to something.

As this suggests, insensitive evidence is often valuable. It is often all we have or can have. But it will not often do what sensitive evidence can do.

The dinosaur case is an instance of inference to the best explanation. The asteroid impact is the best account we can put together of why the non-avian dinosaurs died when they

did. There are many other examples. One more is Wegener's original proposal of continental drift. Alfred Wegener, to simplify the history, noticed that the coastlines of Africa and South America fit together and noticed numerous geological continuities interrupted by oceans. So he hypothesized that continents can drift. Orthodox geologists of his day reacted with "yes, that's nice and neat, but where is the evidence? And how on earth — or in earth — could this happen?" Decades later the study of magnetic alignments and of volcanic rifts suggested that the continents float on plates which move and collide. The geologists said "Oh, Wegener was right, but now we can really confirm it, and see how it happens."

The absence of suggestions about mechanism in the purely geographical considerations creates a barrier to further investigation. They don't tell us which continental movements can be explained by the hypothesis and which cannot. We cannot distinguish between geography that by coincidence is as if produced by continental drift, and geography that really does result from it. Note that some accidental matchings of the geography to the hypothesis might fit it better than geography that is in fact caused by the hypothesized processes. But given just Wegener's considerations, we have no way of investigating which these may be.

The problem here is a more subtle one than in the dinosaur case. Continental drift does cause the shape and distribution of the continents, generally and approximately. And if the precursor continents had possessed very different shapes and locations then the present continents would not have the shapes and locations that they do. So the truth-making facts do cause the observations. But this early version of the theory gives no idea

of the process leading from one to the other. So Wegener was left in much the position of an early microscopist who in the absence of suitable optical knowledge can only say "look into this eyepiece, and you will see marvellous things, for reasons that we do not yet understand." This suggests a somewhat opposite way of putting the point: it would rarely be the case that had Wegener obtained a different, negative, conclusion it would have been because he was dealing with examples where accident rather than continental drift was responsible. (Rarely rather than even a qualified Never because it concerns the character of the evidence rather than the status of the conclusion it supports. There is evidence for false hypotheses, as well as evidence for true but unknown hypotheses.)

positive examples: experiment

These have been cases where evidence is deficient in sensitivity. Contrast them with cases where sensitivity is abundant. Naturally, evidence provided by experiment and measuring apparatus is a source of these. But it is not the only source. Begin with experiment.

One typical biological experiment concerns a species of fish, Bluehead Wrasse, in which the same egg-laying sites are used over generations.[9] The question is whether these sites have intrinsic attractions or advantages, or whether on the other hand there is a "tradition" of using these particular sites which is transmitted from one fish generation to another. The experimenters took fish from two traditional nesting sites on two different reefs and exchanged them. In the unfamiliar reef each used sites which were not the same as those which the resident fish had used. The fish were then returned to their

9 Warner (1988), described in Balcombe (2016).

original reefs; they went back to using the sites that generations of their ancestors had used.

The experiment gives evidence that it is tradition rather than the attributes of a site that influences choice. The familiarity of the reef causes the fish to reuse traditional sites. Moreover we have grounds for knowing that this happens, in that when the reef is not familiar traditional sites are not used. (If not cause then not effect.) Changing familiarity results in changing use. So the evidence is sensitive to the facts that it supports. The return to the traditional sites when the fish are once again in their home reefs reinforces this conclusion, and also helps rule out an alternative explanation of the behaviour in the experiment, that the neglect of the formerly used sites in a new reef is due to the stress of capture and displacement. Confirming both that the fish were influenced by remembered tradition rather than perceived advantage, and that they were not influenced by stress, would have been hard or impossible with passive observation rather than active experiment.

Another example is a classic experiment concerning the replication of DNA. After Watson and Crick discovered the structure of DNA in 1953 and suggested that processes that copy it during cell division might be the basis of heredity, the question of how DNA is copied became important. An experiment by Meselson and Stahl answered some basic questions about how the copying happens. It is an extremely elegant experiment and has been described as "the most beautiful experiment in biology".[10]

10 Meselson and Stahl 1958, Cobb 2015, Franklin 2016).

There were three current plausible suggestions about how the two-stranded DNA molecule is copied. The first was that both strands are independently copied, leading at first to the original molecule and one duplicate and then to duplicates of each of these, and so on. The second was that each strand is copied and then combines with the strand from which it was copied, leading at first to two copies of the original, each of which has one original strand, and then to similar duplication of each of these. (This was Watson and Crick's guess.) The third was that each strand is broken into segments which are then recombined to make double-stranded copies.

The three hypotheses have different predictions about the distribution of atoms and other components from the original DNA molecule in subsequent generations. The first suggestion entails that the atoms of the original molecule stay together and all except for that one are entirely composed of new atoms. The second entails that the original atoms are at first distributed between two copies and later found in exactly two copies. The third entails that the original atoms are at first distributed between two copies and then later scattered between many. Meselson and Stahl saw that which of these consequences is the case could be revealed by exploiting the fact that DNA molecules contain large proportions of nitrogen, which comes in a lighter and a heavier isotope. They grew *E. coli* on a culture that provided them exclusively with the heavier isotope, and then abruptly changed to a culture with the lighter isotope. They extracted DNA both after one cell division and after many, and centrifuged it to reveal the distribution of molecules by weight. (I am omitting many details, inevitably.) The result was that after one generation the molecules were divided by weight into two peaks, and, as the generations went by, the weight appropriate to one original strand was always present though in diminishing

proportions, and moreover the weight appropriate to both strands being copied increased in proportion, in amounts that the second hypothesis predicts. Moreover there were no peaks in the weight distribution at intermediate points. Neither the first nor the third hypothesis predicts this, though the second does.

The experiment provides good evidence for the second hypothesis when it is compared to its rivals. How it does this a topic for chapter 4, but it seems obvious in this case that the evidence does provide this. The duplicating mechanism and the hypothesis are connected in a clear way. The introduction of the isotope causes the separating strands to have different weights which causes them to behave differently when centrifuged which causes the observed lines. Here differential confirmation and sensitivity go together, as each alternative physical possibility would cause an alternative body of evidence.

In this experiment a situation is created in which different hypotheses will have different consequences which can be distinguished quantitatively. There is a deep connection between experimentation and quantitatively formulated hypotheses. If hypotheses can be differentiated in precise terms — is the value of the parameter 0.300 or 0.299? — then fine differences in their consequences can differentiate between the hypotheses, if we can find ways to measure them. This has been an essential feature of science since Galileo. Numerical functions of consequences, test statistics, can allow us to distinguish finely differentiated hypotheses even when our measurements are fairly crude. However we can do this only when we have some assurance that we have blocked unwanted causal processes from interfering.

more carefully

Sensitive evidence is a causal matter, in the general way that includes causation, physical possibility, and the counterfactual (subjunctive) conditional. Anyone discussing such topics needs a flexible and expressive way of making distinctions among them. In chapter 0 I described the apparatus of possible worlds and said that I would use it in spite of the worries one might have.[11] I shall simply use this apparatus in an un-interpreted way, letting readers put their favourite gloss on it. In particular, I shall rely pretty fundamentally on the relation of comparative nearness between possible worlds, as also explained in chapter 0 where I suggest how it can be defined in terms of the counterfactual conditional. Note that as a sometimes inconvenient limiting case this conditional is true when both antecedent and consequent are actually true (since the nearest world to the actual world is itself).

Nearness of worlds can also be used to pin down other "modal" ideas such as what might happen or would be possible if something were to occur, and to explain vacancies and ambiguities in the ordinary use of "possible", "necessary", and the like. (It won't tell us what probability is, though.[12])

Now suppose that we have some evidence, given by a true report *e,* and a hypothesis *H*, and we know that H would cause e if it were true. Take this to mean that in all or most

11 Philosophers have a love/hate relationship with the apparatus of possible worlds. On the one hand it gives a systematic way of connecting and organizing these related concepts, which moreover makes intuitive sense. On the other hand most philosophers find it hard to believe that there are any such things, so they become a handy tool for conceptual purposes, which one hopes to interpret in some more primitive and realistic terms, chosen in accordance with one's philosophical preferences and commitments.

12 This is significant because of the importance of probability in hypothesis testing. I have hopes that ideas about modality and about probability can be used to put a squeeze on one another. But that is another project.

possible worlds from actuality to the nearest world where H is not true, e does not hold.[13] Then when e does occur, an observer can infer that H holds in a range of worlds between actuality and the nearest world where not-H, those where e is true. (Because then given the truth of the causal connection we know that H fails unless e is true.) So H is held in a generally knowledge-like way, with a greater resemblance to knowledge the greater the range of possible worlds, especially those near actuality. It may be that e holds only in actuality. In that case, while the observer may not be irrational in thinking that H , they are making a mistake: although it is true it is not known. This kind of mistake will, given the best intentions and control, often occur, especially on a "realist" understanding of evidence, where the crucial factor is whether the facts are accurately represented because of the interaction between the person's capacity to form beliefs and the specific situation that makes the relevant belief true. It is the inappropriateness of this interaction that results in the true beliefs not being knowledge.

The tighter the similarity to knowledge the greater the range of situations where the hypothesis, as gained by the methods in question, will hold, and therefore, given mild assumptions, the greater the range of situations where conclusions drawn in parallel ways will be true. That is a central importance of knowledge. This range is limited, though, by the range of situations where the evidence will occur. As a result, there are strong reasons to organize experiments around robust ways of producing evidence, those that are effective under variations in the situation. We typically use very robust ways of producing evidence when we make measurements, experiments that are so reliable that

13 This formula resembles both Lewis's and Mackie's accounts of causation, in different ways. It resembles Lewis's core idea (Lewis 1980) in its "if not then not" form, and it resembles Mackie's (1965) in that it embeds a sufficient condition within a necessary condition. Note that Lewis's formula is not suited to apply to non-actual events in its original form, because it entails that every non-actual event causes every actual event. My blowing my nose causes the world not to end. The theme of multiple accounts of causation recurs in chapter 5.

we often do not consider them as experiments. These are themes that will return in the next chapter.

positive examples: non-scientific

Procedures that share basic features of experiment are not confined to science. One finds out if someone is awake by whispering a message; one finds out if there is water in the well by dropping a pebble; one finds out if the enemy is still out there by sticking one's head above the parapet. Moreover these are things that people have always done, sometimes in much the manner of a scientific experiment. There is nothing exclusively scientific about doing something to learn something.

Some knowledge-directed actions are very subtle and push at the limits of action. Among these are social gestures. One raises an eyebrow, looks directly at someone, or smiles in order to discover the other person's reaction. Here is an example from my own life. I was on a committee interviewing for a senior post. A letter for one candidate said that he does not "suffer fools gladly". The committee was uncertain what this familiar phrase might mean this case, so I undertook to find out. At the interview I asked him a really stupid question, confusing two terms. He reacted with ferocity and contempt, settling the issue for us.

The evidence here — the candidate's behaviour — is elicited by a deliberate experimental action. It is also fairly sensitive. The candidate's combination of confidence, intolerance, and tactlessness produces a different behaviour than, say a well-informed desire to help

the other person escape their confusion would have. The interviewer's (my) intervention is designed to produce sensitivity, so that the cause can be read back from the effect.

The interview example is a special case of a strategy that is essential to folk psychology, our everyday understanding of one another. Folk psychology is essential to human life because our mode of operation is based on cooperative activity which is largely not instinctive but thought out task by task. As a result we rely on expectations about what other people will do, and for that matter about what they expect us to do (Morton 2002). But we are remarkably unable to give precise predictions of actions on the basis of other people's information and motives. This is not surprising given that just about any belief and desire is compatible with just about any action, given other beliefs, desires, and other states of mind, including that it would not be crazy to think the person might have. So there is an important question about how we avoid being in the dark about matters that require wariness and trust.

A large part of the answer is that people interact when they engage one another[14]. That is, we routinely do things to test our conjectures about one another. You ask a friend to steady a ladder while you climb it to clear leaves out of a drain trough. Before you even put a foot on the ladder you glance at him to make sure that he is in position and meet his eye to make sure that he is expecting you to go up the ladder now. You look up to where the head of the ladder meets the wall and check that his attention is also focused on this likely problem spot. As you go up, you occasionally stamp one foot or the other in

14 But not the whole answer. In fact, we should not take for granted any easy estimate of how or in what ways we are wrong about one another. One unscientific method that complements the probing that I have described occurs when people in effect agree to adopt certain motives for their interactions. Then they think that they are getting evidence of one another's states of mind but in fact they are guarding against the possibility that the deal has been broken. This is not the occasion to defend this position.

order to confirm that he is holding it steadily and in fact that the vibration is damped by his grip. You do all this automatically, without any reflection or sense of ingeniousness. For that is the way we humans do things together, continually testing our predictions about motive and action.[15]

Thus does experiment-like social intervention buttress folk psychology. Folk psychology operates by a constrained inference to the best explanation, constrained by innate tendencies to understand and interact with one another in particular ways, among other things. The pattern is often found as a way of reducing the dangers of explanation-based inference.[16] We form a hypothesis on the basis of its explanatory power, but often we do not trust what we have conjectured until we have been able to test it with a suitable experiment. One reason why this works is that we can fine-tune experiments so that they rule out alternative explanations. What we are left with is the best available explanation of a larger body of data, which has been deliberately produced so that it will if it can be produced leave much reduced room for alternatives. (There is even less room for alternatives if one is choosing between a fixed very finite set rather than between a hypothesis and all possible alternatives. But then the suitable choice of alternatives is crucial. We will get to that.)

There are broadly analogous patterns with non-deliberate ways of gathering information. Perception provides many examples. Perception uses pretty reliable causal connections between facts and the information, often unconscious, that we receive. These

15 The great variety of experiment-like procedures outside science has a fuzzy borderline with others where one also deliberately produces something although it is not a physical situation. For a controversial example, the use of intuition in analytic philosophy involves careful construction of cases so that they satisfy some criteria while avoiding some traps, particularly rival diagnoses.

16 The travails of the inference to the best explanation, once thought to be the all purpose solution to epistemological problems, are described in Lycan (2005).

connections would not serve their purpose unless they were sensitive: when what you are perceiving changes in perceptible respects then your perception changes. And this is a source of much of our evidence. So what needs to be discussed is not whether we are dealing with a sensitive source of evidence but when the connection is tuneable, whether we can know and control its sensitivity.

Perceptual processes are often tuneable. That is, many details of perception can be adjusted in accordance with perceptual conditions and preliminary output. This can be as simple and automatic as dilating your pupils when the illumination is low, or turning your head so that neither ear points away from the source of sound. There are learned procedures that complement these, such as sniffing deeply to identify a smell when you Ifhave a cold, or for that matter wiping your glasses.

We tune perceptual mechanisms also in response to the plausibility of what they seem to suggest. You think you see a zebra galloping down the railway line, but because this is so unlikely you look again, focusing carefully and shielding your eyes against the light. We can also use similar routines when there is something we want to pay special attention to, or the possibility of something especially interesting. Walking in bear country you check each suspicious movement of large bushes. When you begin to suspect that someone is lying to you, you pay more careful attention to their face and their intonation.

Other connections exploit mechanisms of attention without being prompted by the unexpectedness of an outcome. Some of these, which are labeled "overt" in Mole (2016), are a matter of explicit orientation of body or eyes. When you hear a noise in a

certain direction you look that way. Although in Mole's terminology the attention is overt, it is worth noting that it is governed by unlearned innate mechanisms. Others, revealed by recent cognitive/perceptual psychology, are covert, in that they result from a delicate interplay between the content of levels of perceptual processing and the resources allocated to further processing. For example, Kravitz & Behrman constructed situations in which subjects identified a briefly presented upper case letter on a screen more often when their perception was primed with a prior very brief presentation of the same letter in a related area of the screen, but in lower case.[17] Here a learned perceptual awareness is sensitive to something that would not otherwise be seen, by directing visual resources to it in terms of what is itself an only partially processed stimulus.

Change blindness gives an interesting variety of cases where attention and interest direct perception. People routinely do not perceive "obvious" features of their environment — gorillas crossing basketball courts, changes in colour, the substitution of one person for another, motorcycles directly in their path — when their perception is directed at tracking other things and events.[18] Consider what happens when someone is blind to a feature of their environment that is obvious to someone else. The other person needs only to mention that feature, or to ask a simple question, and resources are redirected so that it suddenly becomes obvious to the first person also.

There are two aspects to the way we respond to unexpected appearances. The first is, as just mentioned, to tune our perceptual mechanisms to get, as it were, a second opinion. The other is to repeat the original perceptual act. We look, turn away, and then in

17 Kravitz & Behrmann (2011)
18 Chabris and Simons 2011)

surprise look again. We do this not just when something appears intrinsically implausible, but when one perception does not cohere with another. This aspect operates even in infants, who stare for longer at events that do not cohere with what they have previously observed.[19] We also follow-up one act of perception with another using a different modality. If we feel a raindrop on the head we hold out a palm to see if its greater sensitivity detects anything. If we hear a crash on the left we turn our eyes and head in that direction. Both repeated acts and successive acts usually give information that checks or complements one another.

So there are two ways in which perception is a particularly reliable source. Both exploit the fact that in perception some environmental events causally produce a change in a person's mind. (a) we can tune — adjust, focus — perceptual channels to make them fit the environment. (b) we can repeat and vary acts of perception to check or confirm what we seem to have learned. These are all connected by being results of our actions to modify the details of the causal chain from event to mind. They are all acting to know.

Even with actions that are not explicitly designed to yield information, we learn whether they succeed or not, and thus get evidence whether actions like this can get results like this, a simple part of the causal structure of things. There is a very fuzzy border between acting to know and acting to accomplish and then incidentally learning something. You can try making a vortex in the pot so that the poached egg does not disintegrate, and this not only gives you a way of making tidy poached eggs but gives you evidence that heat induced solidification can counter the effects of turbulence and centrifugal force. Then you can vary the recipe in many ways, using different implements or different

19 Ballargeon and Spelke (1985)

temperatures or different rotations, and note which ones work. You may do this either to learn how best to poach eggs or to learn about their integrity under various conditions[20].

In a final class of actions directed at knowledge a person influences her own psychology. This can consist simply in making herself pay attention, concentrate hard, or focus on the contents of a difficult book. Mathematicians sometimes take suitable drugs to enhance their creativity and their ability to see complex proofs. (Strong coffee will do, but amphetamines are not unknown for this purpose.) Yet another class consists of choosing suitable sources of information. You go to the best experts, sometimes consulting experts about which other experts to consult. You employ spies, sometimes training them in specific ways. All of these have an obvious experiment-like quality. They all exploit the advantages of sensitive evidence, in particular the ways in which it can be modified and fine-tuned.

evidence, knowledge, agency

My central theme is that when evidence is based on and related to "acting to know" methods it has a number of advantages, which I aim to describe. One might wonder what action has to do with this, since after all a person might wander by accident into a laboratory while an experiment is being conducted, or read a journal article about it, and get important evidence although she is not herself acting in order to know. Is the connection with deliberate action essential? One might also wonder what the advantages are meant to be. Good evidence surely leads often to true belief, but we can get true beliefs without making data, just by analysing it. And how is evidence related to truth

20 This theme was suggested to me by Catherine Elgin.

anyway? There is a connection between these questions.

Elaborate the example of dropping a pebble into a well. You are at the top of a chasm and you want to know how deep it is and whether there is water at the bottom. You notice a pebble moved by the wind roll over the edge, and you realize that it is just the right size and probable weight to bounce off the ledges on the way down and land with an impact you will be able to hear. So you pay attention to it as it falls and eventually you hear a plunk, telling you that there is water down there. And the time of the echo tells you roughly how deep it is. This is sharp observation rather than experiment. You noticed and then watched and then reasoned.

But note two points. First, you have to know what to look and listen for, what to attend to and how to respond to what you pick up. These are actions and you are interacting with what interests you in a way that is shaped by the facts of interest.[21] The analog of experimental apparatus is your own capacity to attend and interpret, which you choose to bring to the situation. Second, there is a price to pay for your limited control of the situation. If the chasm had a different shape or if it was deeper that pebble would not have done the job. So you would have been stymied if things had been just a little different.

All the same, you did learn how deep it was and what was at the bottom. What you ended up with was knowledge. A sign of this is that if there been no water you would not have thought there was. So in nearby but waterless situations you would not have

21 In a loosely similar way mathematical beliefs are shaped by mathematical facts, whose truth is required for the mental processes that lead to the beliefs to work. I suspect that this is one reason that naïve people talk of mathematical intuition and think of it as somewhat like perception.

acquired a false belief. And if you had followed generally the same strategy in slightly different circumstances (longer delay, softer splosh) and got the same answer it would have been because that answer was right. Your conclusion has the character that knowledge is taken to have in contemporary epistemology, consisting in true belief acquired or held for reasons that are linked to the reasons why it is true. Later, I shall have to engage with some of the competing ways of making this rough characterization more precise.

Contrast this with the case in which you deliberately choose a pebble of just the right size and weight to be most likely to give the information you want, and drop it down the chasm. Then you would be more likely to succeed if the chasm were deeper or differently shaped. For you would have chosen the right pebble to do the job under these circumstances. And when you did succeed you would have acquired knowledge, just as in the original case where you were not deliberately choosing and dropping the pebble. So this is a way of acquiring knowledge that will work in a wider range of situations. But it is more broadly reliable in the deliberate choice version than in the passive observation version. So the result is also like knowledge except more so. We could say that it is better knowledge, or more robust.[22] Or we could use a rather restrictive conception of knowledge so that the active version counts as knowledge and the passive version, contrary to the informal labels we are likely to apply to it, does not.

This is the answer to the first question. Acting to know results in more or better knowledge than reasoning from passively obtained observation. This theme will be

22 Speaking of better and worse knowledge is not unprecedented. Of course there are suggestions in ancient philosophy but see also Hetherington (2002).

elaborated. As for the second question, we can see the beginning of an answer. The method of listening for how long the pebble takes to get to the bottom and what sound it makes when it arrives is a good one, assuming that it results in knowledge. And the method of planning which pebble to use and how to direct it is an even better one, assuming that it results in knowledge in a greater variety of circumstances. Good strategies of investigation are those which lead to knowledge; even better strategies lead to more and more secure knowledge. We choose norms of inquiry not because they satisfy abstract criteria of rationality but because they work as our most effective current methods of learning. In these matters the way things actually are and how they might have been play a large role, even if the researchers in question are not aware of these facts. So there is a chance of epistemic failure even when one is proceeding perfectly reasonably and intelligently. One just has to be wrong in certain fatal ways.[23]

One instance of this is particularly important for the present project. Experimentation as currently practised relies profoundly on statistical analysis and statistical inference. The guidelines for doing this correctly would not have been familiar even 200 years ago. They are still being honed and developed, and there are several opposed schools which differ in important ways. So a basic aspect of experimental method is still under construction. It is not as if the reflective common sense of knowledgeable and intelligent people tells us how to do research. Not only do we approach questions of method in the light of what we have accomplished and failed at with previous methods, we have to

23 In this respect the project resembles classic suggestions in Alvin Goldman (1988). See also Elgin (2008). The concept of probability is interesting in this connection. Norm-based accounts of evidence in philosophy and statistics usually depend heavily on the probability of evidence given hypotheses (its likelihood). The relevant probability is nearly always either the researcher's subjective degree belief or their estimate of objective frequencies. The actual frequencies or the tendencies of systems to produce them are rarely involved. So the main traditions here are typically norm-based rather than knowledge-imitating. But it does depend on one's attitude to probability.

struggle to discern accomplishment and failure.[24] This is so if we accept that it is very often harder to know what one knows about a topic than it is to know truths concerning the topic. Knowing that one knows is typically harder than knowing.[25] As a result, the material for reflection on method is typically rarer and more uncertain than the results of method. This is so when we are discussing experiment: basic aspects of experimentation can only be settled by reflection on the success or failure of particular experiments.

24 The introduction to Williamson (2000) is relevant here. Morton (2012c) and Morton (2013) argue for a thoroughgoing symmetry between knowledge and accomplishment. Morton (2014a) makes the link with experiment.

25 Williamson 2000 ch 5 and appendix 2

chapter 2: experiments as causal processes

The experiments that provide evidence are real physical processes, and the information they provide is made possible by their causal structures.[26] Very often the information that they give itself concerns causal structure, a collection of linked causal relations between events or physical quantities. One question that this chapter begins to address is how the causal structure within an experiment gives information about its target. This chapter begins to sharpen a picture of the particular kind of evidence that scientific experiments, and their everyday analogues and precursors, provide. The first task is to be explicit about experiments as causal processes.

cause, counterfactuals, evidence

The account of experiment and of evidence generally in this book is thoroughly causal. I take this label in a very general way, though, so that it includes counterfactual (subjunctive) conditionals and a wide variety of causal relations. In chapter 5 I will discuss how evidence can distinguish between these various relations. But a description of experiments in terms of counterfactuals will give a sense of how we can connect their structure with the character of their evidence.

We are interested two incompatible hypothesis, H_1 and H_2. (Both might be false, but only one can be true.) There is possible observable evidence e_i ($i \varepsilon\{1, 2\}$) relevant to the choice between them in that if H_1 were true e_1 would be observed under normal or frequent background conditions, and if H_2 were true e_2 would be observed. Just observing

26 Thought experiments are another matter, but I am not going to discuss them.

e_1 or e_2 will not tell us much. Both hypotheses may be false in the actual world and e_i true there, although the nearest world in which either holds may be one with the corresponding observation. (If a civilization on Mars had built canals, we would observe them from Earth; if a civilization on Mars had refrained from building canals, we would observe no canals from Earth; we observe no canals from earth; therefore there was a non-canal-building civilization on Mars? A lot of sophistication has gone into filling the gap here.) So instead of tangling with all the competing complications and possibilities we set up an experiment. We create a situation S, often involving a process producing e_1 or e_2 set off by a deliberate triggering action, where given S "if e_1 then (if H_1 or H_2 then H_1)" and "if e_2 then (if H_1 or H_2 then H_2)" are true. I refer to these as *reversal conditions*. We will see more of them. (All the conditionals I use will be subjunctive/counterfactual, unless otherwise indicated.) Then when e_1 is observed we can conclude not the impossibly simple H_1 — science is not that trivial — but the useful and informative information that H_1 is true in a nearer world than any where H_2 holds. (It is more nearly true; it would take less of a variation, if any, on actuality to make it true; it would have taken a smaller and later variation, if any, from the actual development of the universe to produce it. These are obviously not all equivalent, but they will do for the moment.) The effect of S is to reverse the dependence between the hypotheses and the evidence.

Measurements are very simple experiments (see below). They illustrate reversal principles and give some confidence that reversal is possible. If we spill a quantity of mercury on the table its changes in size will not tell us much about the atmospheric pressure. There are too many possible influences. But if we surround the same quantity with a carefully designed apparatus and calibrate it suitably it will be a fairly reliable

indicator of the pressure. Many of the irrelevant influences will have been excluded or minimized. Each marking on the dial is a possible outcome corresponding to a possible pressure, and the physical setup has transformed "if pressure then marking" to "if designed and calibrated then (if marking then pressure)". It is because of this that it can be informative.

We need pairs of conditionals we are comfortable with saying what would happen if an experiment were carried out in particular conditions to compare a particular pair of hypotheses. And we need them to say what would have resulted had the facts been one way rather than another. I suspect that the ideal tool here would be a contrastive counterfactual "if a_1 occurred rather than a_2 then c_1 would occur rather than c_2" with a dedicated semantics in terms of differences between worlds that highlight the differences between a_1 and a_2. Then there would be a clearer incommensurability between results of different experiments. But developing this would risk burying the message in the machinery, so instead I shall use pairs of conventional Lewis/Stalnacker conditionals "if a_1 then c_1", "if a_2 then c_2", with emphasis on pairs "if t when H_1 then e", "if t when H_2 then f". (f will often be simply the negation of e, but sometimes it will be a more subtle alternative to e.) "When $H_{1(2)}$" alludes to an inevitable complication. Sometimes one of the two hypotheses will be true and a law of nature. Then a naïve antecedent "if H…" is problematic; H is true in many or all causally possible worlds. Also, often combined with this, an alternative to H may be causally impossible; counterfactuals with this alternative as antecedent are equally problematic. Though I shall rarely indicate this explicitly, I shall treat such counterfactuals as if they were bundles or enormous conjunctions of conditionals stating what would happen in particular cases of the necessary or impossible

hypotheses. ("If energy were not conserved…" may be of dubious intelligibility, but "if the combined kinetic energy of the particles were e …" when it is in fact e' ≠ e, is a lot better behaved.)

Here is a simple argument to the conclusion that experiments as described in this chapter lead to such reversal principles. (Too simple perhaps. Complications in a moment.) Suppose that either H_1 or H_2 is true. Then when the experiment is performed either e_1 or e_2 will result, but not both since they are incompatible. Now suppose that it is e_1 that happens. e_2 will not occur. But if H_2 had occurred e_2 would have been observed, so H_2 is not true. So if exactly one of them is true it is H_1. It remains possible that they are both false, but at any rate we have that if one of them is true it is H_1. Similarly on observing e_2 we can conclude that if one of them is true it is H_2.

The fly in the ointment is knowing that the requirements for a reversal principle are met. The history of science is full of experiments that were thought at the time to be crucial, and establish which of two hypotheses were indicated. Even taking into account that the aim here is not to determine which is true, holding in actuality, but which is true in a situation nearer to actuality, there are many ways to be mistaken. One requirement is that if a hypothesis is true the experiment give a particular outcome. We can very rarely deduce a description of the outcome, from a statement of the hypothesis alone. So this requirement is counterfactual: if H *were* true… . But there are many unknown counterfactual conditionals, and we often think one is true when it is not. A similar point goes for knowing that the outcomes associated with the two hypotheses are incompatible. A form of both of these, which is so frequent as to be practically universal,

occurs when the relevant outcomes cannot be observed in a finite time. The practically universal form occurs most importantly when the evidence is statistical, and no single experiment will definitively give answers. (Series of trials are discussed at several places in later chapters. The very fact that different trials can give opposing indications shows that there is a lot more to discuss.) An infinite series of trials or a sample consisting of the whole population would be enough, but the central need for statistics comes because these are practically impossible. Links between statistical hypothesis testing and the framework here are in chapter four.

Another basic gap to fill concerns ignorance and error about objective evidence, as a result largely of our shaky grasp of which counterfactuals are true. Epistemic agents are frequently wrong about the strength of their evidence, and regularly have more or less evidence than they think they do. The phrase "on observing e_2 we can conclude ..." above could do with more explanation.

So the connection between experiment and objective or sensitive evidence is not yet fitted to a lot of what we do when finding support for a hypothesis. But it is now clear that there are links at a basic and general level, comparing hypotheses in terms of which is true under a greater perturbation of however things actually are.[27] Honing the connection is a recurrent task in what follows.

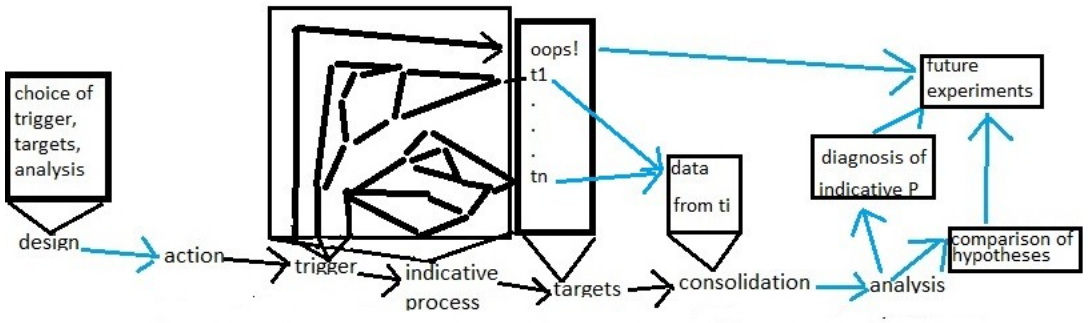
informative processes

27 This is loosely related to what Woodward (2006) calls "sensitive causation".

An experimenter hopes to learn about a topic. She may well have a specific hypothesis in mind, though purely exploratory experiments are far from unknown. So she contrives a situation where a process, the "*indicative process*", occurs. She may have to set this process going, or it may happen spontaneously. A trigger for it will often involve putting the value of some quantity in a particular range. There is usually a range of anticipated consequences c_1,..., c_n, of the process — the targets — which also often consist in quantities lying within ranges. The hope is that the consequence that actually results, for this combination of factors, will feed an inference revealing something about the topic.

This general description applies to many interactions probing for the origins of a phenomenon. There are standard rules for doing this, discussed in the next chapter. The question now is how the familiar forms of experiment fit the pattern of the previous section. The familiar routines of experiment can be spelled out in diagrams as follows.

The diagram includes a "transducer" step where the outcomes of a number of occurrences of the indicative process are transformed into a single output. This is to accommodate the inescapable fact that the data can usually only be interpreted in bulk. Many individual items will be misleading or irrelevant in isolation so that it is data sets rather than data items that constitute evidence.[28] Then a standard causal pattern of experiments can be represented as follows.



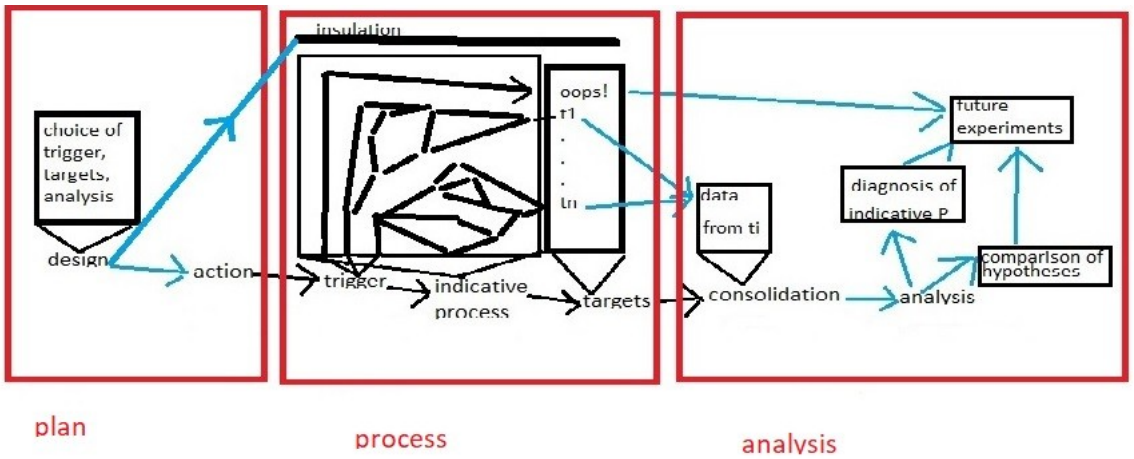
ıd cases where the
2011).

The black arrows represent causal processes that actually occur, whether or not the experimenter has anticipated them. The blue arrows represent thoughts and intentions of the experimenter. They may contain mistakes or be based on false assumptions. The pattern can be summed up as design leading to a trigger leading to a process resulting in targets whose information is consolidated prompting an analysis. Each of these six stages has a standard content, as illustrated in the boxes. Call this the six box pattern, for want of a better name.

Although the arrows are causal, the events on their left are rarely, perhaps never, the only causes of the events on their right. And in particular there are nearly always other causes that do not pass through the events on the left. If they have a large influence then the final inference is muted or complicated. (The initial design usually requires them to be minimized.) So ways of preventing outside influences are usually put in place. I shall refer to all manner of preventions as "insulation". Insulation is typically selective: it depends on what we want to be taking account of in the hypotheses we are comparing. So it is a matter of planning at least as much as of physical operation. Building this into the diagram we can impose a simple schema on all the boxes. One begins forming a plan, then one carries it out, and then one analyzes the results. This gives us what I shall call the *three box model*. The middle box, process, is importantly different from the other two. It contains what actually happens, whether or not it is expected or understood. The independence of what happens at this stage from the other two is important.

The now simpler structure, still with quite a lot of detail is as follows:

The



This is the pattern of many classic experiments in the history of science, and of much contemporary research. We are used to taking it as a source of reliable evidence. For a clue about why such pattern might be the source of such as results return to the reversal conditions of the previous section. These state that the evidence gives one hypothesis a preferential status to the other, in terms of the remoteness of the situations where it is true. So why might we expect the structure of an experiment to be correlated with evidence having this power?

It is tempting to think that the main factor must be the insulation. It allows some control of what influences the process and its outcomes. But this cannot be more than a small ingredient in many cases. Insulation is a crude measure; it obstructs whole categories of influences. But we usually need to tune the influences to fit a particular hypothesis or hypotheses. We might be interested in whether one virus *in the absence of* a particular other is correlated with a symptom. Or more subtly whether radiation *between one level*

and another, just slightly different, can induce a mutation. And in fact there do not seem to be many examples among the classic experiments in the history of science where it is insulation alone that makes the experiment work. A general way of putting the point is that evidence supports or undermines hypotheses which have many logical forms, while insulation can at most give a list of factors allowed and factors obstructed. This is also the reason why the language of confounders is not very helpful.

We do control what factors influence indicative processes. And we can do this in a structured way, often without realizing exactly what we are doing. One way we often do it is by combining a fairly broad-spectrum insulation with a much more specific trigger. We tune the trigger to provide the combination of factors that we want and we rely on the insulation to ensure that a range of factors beyond this combination is not affecting the results. There are many examples in science. One very clear one is Frances Arnold's Nobel prize-winning work on the production of enzymes by directed evolution.[29] Arnold manufactured organisms which were exposed to conditions leading them to produce enzymes with desired properties, such as the synthesis of biofuels. In a preliminary stage the organisms are created by a very deliberate evolutionary process. Then in the part of the experiment that illustrates my point these deliberately construed agents are used to trigger a chemical process which, without special insulation for the hypotheses in question, produces the intended targets. This process is shielded from outside influences in standard ways which are not particularly discriminating. The result is experimental evidence for, among other things, the complex hypothesis that there is an enzyme that facilitates the reaction in question. Once one sees the pattern one can find numerous other cases.

29 AAAS (2019)

The force of the evidence, to the extent that it depends on whether a specific combination of factors is responsible for the target achieved, depends on what actually happens during the experiment. This can often reduce to a question of whether it would have proceeded differently under various other conditions. Experimenters can therefore be mistaken not simply about which hypothesis is true but about how well the result of an experiment supports one hypothesis over another. This shapes one's picture of scientific rationality, but it also has an impact on practical issues of method. In the case I have just described, for example, we have a two-stage experiment. First one produces the organism in question, and it is necessary to confirm that the procedure will do this reliably, and then one uses the organism to catalyze a chemical process, and for this it is necessary to confirm that organisms producing this enzyme will facilitate the required catalysis. But this second part of the experiment presupposes that the hypothesis in the first part has been confirmed. A perfectly competent scientist might conduct the second part of the experiment thinking wrongly that the first part had been a success. While this might simply be a matter of bad luck, all the same the overall project would be a failure.

The result is that there can be conditions in the performance and surroundings of an experiment that satisfy a reversal condition. Their effect is that an evidential event that in other circumstances will be produced if one hypothesis rather than a rival is true will under the specific experimental conditions *only* occur if one rather than the other holds. An enzyme which would normally be produced by a biological process if it exists but which could also be produced by other processes is under the conditions of an experiment only possible by that particular process, thus giving evidence that the

process does occur. (“Only possible” meaning that this is what it requires except in really exotic scenarios.) There is no guarantee that such conditions can always be found for a pair of hypotheses. Our knowledge is in the hands of nature, but our ingenuity allows us many ways of persuading nature to cooperate. Even when we can achieve it the reversal is usually imperfect. If one hypothesis rather than the other is true then it is likely that one event rather than its contrast will be produced, and given that one has been produced when the other was conceivable its origins are likely to lie in the truth of the one hypothesis rather than the other. These inevitable probabilistic weakenings are a topic of chapter 4.

Sometimes the trigger is absence of an intervention, and then the insulation is important for indicating what will happen in an unperturbed state, typically one that is rare in the uncontrolled world. An example where nothing will happen and the trigger is a non-trigger is given by Pasteur's experiment undermining the idea of spontaneous generation. Boiled meat broth was left to cool under S-shaped tubes which allowed air but frustrated spores, resulting in no worms, insects or whatever. The hypothesis that organic matter left alone will not generate animal life is supported, and again the wide-spectrum insulation allows a transition from what may be expected if the hypothesis is true to what is suggested if this expectation is fulfilled. (Although one that makes a fair number of assumptions and needs further evidence before it is at all conclusive.)

A tricky issue is the uniqueness of the articulation. Could we make two quite different diagrams of the same experiment? I am not sure how to tackle this question, but one obvious concern is the distinction between trigger and insulation. What on one

articulation might be initiating a trigger could on another articulation be preventing interference or drowning out of a causal factor that would occur anyway. The application of a source of radioactivity might be an example. Intuitively, insulation is double prevention: preventing unintended factors from preventing the trigger from having its effects or from their being detectable. And one distinctive of prevention is that it presupposes a deeper cause whose operation it hinders, one that will operate in a wider range of situations. (The depth of causes is discussed in chapter 5.)

experiments within experiments, measurement

Each of the three boxes can itself contain a whole experiment, itself dividable into parts. Two frequent occasions for this are in preparation of materials and in measurement. It is often not obvious that the materials for an experiment are as it requires. Then a subsidiary experiment can check. It can be as simple as sampling from a random selection of candidate materials, or itself more like a whole insulation/trigger/process experiment. One example, from Arnold's work on targeted evolution, has been mentioned. Similar procedures are fairly common in biology. Another, also a Nobel-prize-winning series of biological experiments, is Ohsumi's work on autophagy, which involved developing a yeast where processes of recycling used cellular materials proceed slowly enough that they can be studied.[30]

The resemblance between experimental apparatus and observational instruments is vivid with microscopes. There we often cause changes in the object we are observing. A simple

30 Ohsumi and others 1992. Another important resemblance between experiment and measurement is the need for statistical techniques to separate signal from noise (Bland and Altman (1996), Hawkins (2014)).

case is the use of dyes with optical microscopes in biology; a suitable stain brings out the features of interest.[31] The changes in the object may not be trivial, and in fact we often have to take care not to destroy it.[32]) The line between observation and measurement is also pretty vague. (Is a theodolite a kind of telescope or a kind of protractor?) So sometimes it the purposes of a procedure are more important than in its structure in classifying it as experiment, observation, or measurement. Many varied enterprises of changing things in order to know more about them use the three box patterns of causation, for they are effective ways of making forceful evidence appear.

(Values of fundamental constants are routinely obtained in this way, even when there are conflicting theories about their values and how the apparatus works (Ekstrom and Wineland 1980).)

and outside science

These strategies for sharpening experimental evidence are generally similar to ways that we get evidence in everyday life to focus more precisely. Two examples illustrate this, and hint at what may be special about the form they take in science.

First consider selective attention, where one ignores information that is not elicited by a particular controlled action. A blind person is pretty good at navigating through a complex environment guided by the echoes of ambient noises. There are limits to how well this works, though. So instead very often this person taps on the ground with a cane and guides herself by the direction and timing of the echoes. She knows where the sounds come from and which ones are echoes of which, and this gives her a less fallible

31 Alirio Rosales showed me the importance of this simple fact.

32 There are fewer cases with telescopes in astronomy than with microscopes in biology, for obvious reasons. But the point is not completely inapplicable: there have been observations of the moon which involve shining lasers at it from earth.

grasp of where obstacles are. She has gone from "if there is a wall there then there will be an echo from this direction sounding like that" to "if there is an echo sounding like that from this direction then most likely there is a wall there". But to do this she has to ignore the competing flurry of echo-information from other sources. (Note that this is also an example of the common ground between observation and experiment.[33])

Next consider informal randomization. A person is curious whether a switch controls a distant light. It would not be the only switch connected to the light, and it is possible that the power supply is intermittent, affecting both the light and the unreliable switch. So he flips the switch in an "unnatural" way, perhaps beginning *Paradise Lost* in Morse code. If the light then signals that very sequence back to him, even with a small degree of fuzziness, he can be pretty sure that the pattern of switch flipping lies behind the pattern of illumination. Again there is a reversal. He can go from "the way the switch is flipped is a possible cause of the on-and-off of the light" to "the overwhelmingly most likely cause of the light's turning on and off is the action of the switch". (There is also a causal dimension here, which I return to in chapter 5.)

For a third example consider everyday psychological attribution. As I noted in the first chapter, we often do this by watching for people's reactions to our initiatives. I suggested that this was one source of our acuteness about one another's minds. The reversal aspect can be found here too, but it is a different feature that I want to emphasize. We can base our thoughts about others on their facial expressions, body language, and general demeanour, which can have many sources besides their reactions to our

33 Another way that observation fades into experiment, especially if we see the continuities between everyday perception and instrument-mediated scientific observation (Brown 1987).

probings. We can also focus on what they say, and in particular on their choice of expression from a limited range in response to a demanding instruction. "Would you describe your attitude as resentful or angry?" "Rate the candidate's suitability for graduate work on the following five-point scale." This allows us to choose between more precise alternatives. There are two advantages. The first is the smaller number of possible reasons why one of these specific alternatives would be true, at any rate given background assumptions and attitudes. This makes reversal easier, and thus evidence more definite. The second advantage is just the more specific nature of the hypotheses. Since they are attributions of attitudes to particular propositions each of them can be contrasted with yet further possibilities and these afford yet further possible confirmations or tests. Thus in the long run we have a better chance of winnowing out an accurate picture of a person's states of mind.

Three things — precision of hypothesis, fruitfulness, and force of evidence that may be gained from a successful experiment — go together in these everyday cases. They are generally associated, and this has broad epistemic consequences. That is the topic of the final section of this chapter.

forks and circles

Many general theories contain undefined parameters which take different values in different applications. The downward acceleration of a falling body depends on its mass and that of the planet it is falling towards. Weighing Earth or Mars would be tricky but measuring acceleration is somewhat easier, so we can estimate these values in familiar

ways. One standard technique is to try out successively finer pairs of values, zeroing in on the final estimate. We could first compare the hypothesis that bodies on earth descend at (roughly) $10m/s^2$ against the hypothesis of $5m/s^2$, then 9.9 against 7, and so on, getting nearer and nearer to 9.80665. (With this well behind us we can compare the values at different points on the Earth's surface.) Values of fundamental constants are routinely obtained in this way, even when there are conflicting theories about their values and how the apparatus works.[34] The procedure here is midway between experiment and measurement.

The parameter-fixing example illustrates what can be done. We know that the equation giving acceleration in terms of mass is essential to explaining the data about a particular falling body because when we vary the values in this equation we no longer connect with the data. If they had been different, as they could have been on a different planet or with a slightly different history on this one, the events actually produced in the experiment would not have occurred. More generally, an item of evidence supports a component of a larger theory when that item would not have occurred if some variant rather than that component had been true. Conditionals like this rarely hold except under experimental conditions.

The quantitative emphasis in scientific theory facilitates this technique, by giving opportunities for finely differentiated hypotheses and correspondingly precise measurements. Using varied propositions as objects of mental states does something similar in everyday thinking about mind.[35] The technique is generally reversal friendly

34 Ekstrom and Wineland (1980).

35 Numbers and propositions are not the only two possibilities. Any domain with a finely differentiated structure would do.

and can boost evidential support. The quantitative focus and the empirical focus are symbiotic. But there is a link here with another desirable feature of theories.

Often different parts of a theory do the work in explaining different classes of phenomena. And theories quite often have redundant parts that, intuitively, are not earning their epistemic keep. It would be desirable to target evidence so that it is clearer which parts of a theory it applies to. But this is notoriously difficult with accounts of evidence centred on explanatory force. Crudely, a bloated theory explains as much as it is leaner functional parts.[36] But an account of experimental evidence based on the evidence-providing events that would occur if one hypothesis rather than another were true opens up more possibilities.

The possibility of separating evidence for different parts of a theory is never guaranteed. It depends on whether there are experiments which will in fact have suitable outcomes, and thus on the causal facts as they are. This goes some way to explaining the conflict between intuition and doctrine. Since we can isolate parts of many quantitative theories in experimental science in terms of their relevance to particular bodies of data, we think of this as a feature of theories, and of confirmation in general. Moreover it is a methodologically nice property for theories to have. But attempts to find a feature of theories in general that accounts for this property fail, because there is no such feature. It depends on what the physical facts are and what experiments we are able to perform.

36 This is the issue which Glymour's bootstrap method, a prescient rebellion against the prevailing epistemic holism of its day, is meant to address. Glymour (1980), Christensen (1983).

The capacity to combine as well as divide theories is important. It provides some protection against dangers of circularity. Constructing an experiment usually relies on previous theories, which may be well confirmed standard doctrine or improvised rough modelling. When these theories are false the results of the experiment may be misleading. For a trivial example suppose that we assume that the material of a metre stick does not expand as it becomes hotter. Then we experimentally heat a range of substances to see how they are affected by temperature, measuring them with metre sticks. If our assumption about the metre stick is wrong then the conclusions about these substances are likely to be erroneous. A slight error in the assumption could lead to large errors in the conclusions. There is a practical problem in experimental design here, linked to an abstract sceptical worry about the significance of evidence.[37]

But a standard experimental technique gets around this problem. We can base experiments on different theories. We can measure by purely optical measurements how objects expand and change their shapes when they are heated so that nothing like the yardstick is involved. Or we can use different materials with the standard ways of measuring. The result is more convincing the stronger the evidence for the theory we are now using, and so the force of this point about joining theories together depends on the previous point about taking them apart. And we can do it only when nature lets us. But the result is progress. When true assumptions are used to construct and interpret experiments where the phenomenon produced are in fact the results of the experimental processes then well confirmed theories will often be true. That broad generalization relies

37 The worry is most familiar to epistemologists as a point about disconfirmation, in the guise of Quine's use of Duhem's observation that when evidence seems to undermine a theory we can usually instead take it as undermining some ancillary assumption. If this point is combined with a resistance to allowing evidence to target parts of complex theories, then the result is a widespread indeterminacy. We seem able to react to any evidence in an enormous number of ways without incoherence.

on our having enough true alternatives to a false theory, and of course we have no guarantee about the truth value of any of them. But as long as we can combine parts of larger theories into suitable experiment-supporting complexes and as long as we vary our replications of experiments, they will buffer us against self-confirming circularity. When ingenuity is there and the facts are friendly, that is. (Issues about singularity and differential confirmation of parts of theories will return in the final chapter,7.)

The characteristically experimental force of evidence is thus knowledge-like in its lack of iteration: having strong evidence for a hypothesis does not mean having strong evidence that one's evidence is strong. The facts about the subject matter and the facts about one's relation to it are independent. A suitably structured experiment comparing a hypothesis to a suitable alternative it can result in evidence that occurs only in situations like those where it is true. Moreover this evidence will often discriminate between large theories and their components, and guide further experiments, leading to an accumulating body of knowledge. Those are its virtues, when we are lucky enough to have it.

chapter 3: the rules of experiment and the success of inquiry

norms of experiment

There are norms of experimentation, criteria for a well-conducted experiment. These vary a little from one discipline to another, but there is a common core which is essential to scientific practice, and which researchers violate at their peril. The more closely the production of data and its analysis conforms to them, the better the evidence for a hypothesis is regularly taken to be. This is the first of two chapters about standards of the force of evidence internal to science, discussing them from an epistemic point of view.

The message for epistemologists in this chapter is that the reasons why the norms of experimentation are generally effective suggest general strategies for acquiring knowledge, and the way they are generally formulated makes it harder to see some interesting issues that arise from them. The message for philosophers of science is that these norms are not entirely unproblematic, and raise interesting issues, about causation and its discovery and also about the interpretation of a hypothesis-as-confirmed.

The existence and near universal acceptance of these norms raises a number of questions. They include:

In what sense is evidence better or stronger if it conforms to the norms?

What is their rationale: adherence to them brings what advantages or forestalls what dangers?

> How embedded in a particular scientific tradition are they? What precedents or analogues are there in non-and pre-scientific practice?

A naive preliminary question is why we need such norms at all, beyond basic principles of honest data record keeping and enough statistics to support simple inductive reasoning. Suppose that we are investigating the heights of people in Vancouver in 2019. We take a sample and measure them, and then accept or reject a claim about the average height of all Vancouverites. Questions about how to get to that acceptance or rejection are deferred until the next chapter. (They are pretty familiar questions, but I put a twist on them.) We do this without control groups, randomization, and the rest. What could be wrong about the result? It seems to be the best we can get as a stab at the average height of people in Vancouver (and similarly for other aspects of Vancouver heights). After all, factors that affect the sample are likely to affect the larger population, and if the sample is not typical of human heights in general, or people in affluent countries in the early twenty-first century, then that particular population, Vancouverites, is not going to be typical either.

There are two things wrong with this reasoning. The first is a matter of suitability. Is a sample going to tell us anything about this population? (Or about the heights of humans in general.)? Suppose that people get to live in Vancouver as the result of some completely random process. Permission to live there might be the prize in a lottery won by some 600,000 people. And that is the only way someone gets to live there. The result is like an accidental generalization from 1960s epistemology (coins in someone's pocket, New Jersey school board presidents). Or consider two variables defined as functions of

completely random variables (results of two lotteries, perhaps), which for no reason happen largely to coincide. In such cases the sample will tell us nothing about the wider population (either those who happen to live in the city at the moment or the long-term population). It is not the kind of fact that can be projected.[38]

(It is customary to contrast two categories of inference. On the one hand there is inference to a hypothesis which is noncommittal about whether some of the correlated terms cause others or whether some of the correlations are effects of common causes. And on the other hand there is inference to hypotheses about which factors are causes of which others. This second kind is standardly called "causal inference". However there is a sense in which all inference is causal inference. If there is no causal background at all for a generalization then we can only come to know it by knowing each of its instances separately. So when we make a standard inference to a generalization we are presupposing that it is backed up by some systematic causal reasons. These can be stronger or weaker, and there are lots of varieties, tending to pure coincidence at one extreme and direct causal connection at the other, from an easily-perturbed exception-prone patterned case-by-case to a causally and counterfactually robust fundamental law.[39] I have more to say about this later (chapter five). But one is always reasoning to a conclusion about causal organization.)

The second problem with the reasoning is a matter of intended application. Contrast two uses for the eventual conclusion about the whole Vancouver population. Suppose that we are going to use it to order clothes that will outfit everyone in town for a special

38 A curious feature is that most patterns are like this, although examples do not occur readily to us. (Similar to the surprising fact that most real numbers are irrational, though most of us can only name a few.)

39 Johnson (1991), Sober (2001)

occasion. Then we will want to know (among other things) the average height of these particular people as they are now. But suppose on the other hand that we are going to use it to design new buses which will have a service life of thirty years. Then we will want to know the average height to expect of people subject to the conditions found in Vancouver. It is easy to think of further uses and corresponding reference classes. The important point is that the first class projects from the sample to a set of actual people, while the second class projects to potential people and actual people as they may develop.

These two issues can coincide. Suppose for example that we are studying animals that do not form a biological kind. They are no species or subspecies, but are grouped together in everyday language and thinking. ("Bugs" as including animals as different as spiders and grasshoppers, "geese" as including some kinds of ducks.) Then samples from the commonsense kind will give very limited information about the disparate biological kinds, and samples confined to one of these biological kinds may not answer the questions we are asking about the whole intuitive kind. (If we want to know mechanisms of reproduction we should stick to the biology; if we want to know how to raise and manage them our questions may be best addressed to the intuitive kind.) There will be a link between the particular generalizations aimed at and the purposes we want them for which will shape the appropriate causal origins of the samples and corresponding populations we take account of. So inappropriate sampling, inappropriate for the purpose, should be avoided. This is often described in terms of a vague rhetoric of avoiding “confounders”, unwanted factors that can obscure the conclusion. But what needs to be avoided will vary depending on the full conclusion at stake. Much of this

chapter takes the standard rhetoric and specifies the influences in more informative ways, more sensitive to the causal origins intended..

One reason for the language of confounders is thinking that the aim is simply to achieve truth. But as many writers have pointed out we want not only to believe truths but to avoid falsehoods.[40] And the concern could be acquiring a true belief for oneself, acquiring a true belief for a community of researchers in the field, discerning a promising line of inquiry for oneself or a larger group, or coming to know why a phenomenon occurs. It is not at all obvious that all of these are promoted by the same measures and the same understanding of evidential force. Neither is it obvious that different stages of investigation — such as spotting promising topics and hypotheses, deciding between rival promising possibilities, and replicating apparently successful experiments — call for the same safeguards. For the rest of this chapter I shall go through the standard interference-protection norms of experimentation, suggesting which kinds of protection against which kinds of interference and experimental procedure are helpful for which purposes.

First, more about encouraging truth and avoiding falsehood. There are two standard frameworks, one more common among epistemologists and one more common among statisticians and scientists. The epistemologists contrast error avoidance, doing what works best to make it unlikely that one ends up with a false conclusion, with ignorance avoidance, doing what works best to make it unlikely that one ends up without a true (and useful or informative) conclusion. Avoiding error with no concern about relief from

40 Perhaps the first person to point this out was William James. See page 30 of James (1897). More recent expositions are chapter 7 of Levi (1974), and more accessibly chapter 1 of Goldman (1988).

ignorance is trivial: one accepts none but the most certain conclusions. But this is no way to understand anything difficult. Avoiding ignorance with no concern about error is also in a way trivial: one accepts conclusions as soon as they provide some sort of hold on the phenomena, with little regard to how they fit together or whether some of them may be wrong. The result would be that true accounts of the topic hide scattered among the varied and ill sorted parts of the resulting theories. In statistics much the same work is done by the distinction between type I and type II errors. Type I is haste, changing one's view too readily, resulting in a false belief. Type II is over-conservatism, sticking with a prior account even when this would be to retain falsity (so missing an opportunity for truth). The presupposition is that there are two hypotheses, the null or default hypothesis and the alternative or speculative hypothesis, and there is a source of probabilities for the data that will guide the choice of hypothesis. It is also assumed that one is choosing between on the one hand retaining the null, which can sometimes be so vacuous as forming no explanation of the data at all, or on the other hand abandoning it to accept the substantive alternative hypothesis.

Though the general idea behind both the epistemological and the statistical version is the same there are significant differences between them. The standard statistical version is thoroughly probabilistic, and gets its probabilities from the two hypotheses concerned, as described in the next chapter. The crucial probabilities are those of wrongly retaining the null hypothesis and of wrongly accepting the alternative; they represent the chances of making these mistakes with respect to that particular pair of hypotheses given that particular evidence. These are probabilities as signs of justified confidence. On the epistemic version the concern is with the possibilities for truth or falsity in using the

method that that particular belief at that moment. This can have a broad focus on the chance of truth or falsity in the longer run as in many analyses of justified belief, or a narrower focus on the range of possible situations where the acquisition would have been successful, as in most analyses of knowledge. My emphasis will be on the latter. The probabilities are signs of reliability. They indicate how strong a tendency to truth or falsehood an epistemic practice has. This gives them a favour of causal possibility. The causal significance of the probabilities involved will turn out to be a mark of differences between norms of experimentation.

Thinking statistically about probability in connection with someone's coming to a conclusion, then, directs attention to the probability of truth among other conclusions obtained "in the same way". Thinking epistemically directs attention instead to the possibility of truth or falsity of that very conclusion for that person then. Both of these combinations — probability with method and possibility with situation — are important. We have cause for worry if either is attributed in a way that does not serve its purpose in the method where it is used. But they are different, and combining them is a delicate matter. I do not tackle this until the next chapter. My concern now is with the occasion-to-occasion possibilities of truth and falsehood for conclusions that respect or violate individual standard norms.

ten commandments for researchers

There are standard rules or norms for researchers. Their origins, often reflected in the terminology, are largely in agricultural and medical research, although they are now

spreading throughout science, even into particle physics. They come with standard rationales, and it is these that need attention here. Each in turn.

(1) *plan* experiments before they are carried out, including choosing the size of the sample and the way it is selected, and the way that the data is going to be analysed. The point of this that there may be no profit from the experiment unless it has a particular evidential force, usually requiring a particular sample size, method of analysis, and so on. It is a waste of effort to do the work unless the results are worth having. This presupposes that the force of the results depends not just on the content of the evidence but also the way in which it was produced. It depends also on the purpose of the experiment, exploration, confirmation, replication, or whatever. One basic and often neglected task is to be explicit about this and plan accordingly.

A contentious issue about planning concerns advance decisions about when to end a trial. The majority of statisticians warn against deciding only during an experiment the point when enough evidence has been collected. The reasons are clear enough: if we decide to call a coin biased the moment that there are many more heads than tails or vice versa, and to continue the experiment until that point is reached, then we will almost always conclude that a coin is biased, even when it is not. Better, according to orthodoxy, to decide in advance how many trials to make and stop at that point. Bayesians sometimes disagree with this advice, and the technique of sequential testing provides ways around it in some circumstances. So there is a connection with issues about the sources of probability. But on nearly all philosophical accounts of probability there are probability

assignments one is confident in making and those that are more conjectural.[41] If the point of the experiment is to uncover possibilities that are worth further investigation then it would be defensible to use somewhat conjectural probabilities, perhaps derived from a possible but not inevitable model of the situation, and to cease experimenting when an interesting pattern emerges. (The coin lands heads five times in a row. It is reasonable to suspect it might be biased, though this is a long way from conclusive evidence.) On the other hand if the point is to make a more definite decision about the truth of a hypothesis then this policy would very often be a bad one.

The relatively uncontentious distinction between suggesting research and accepting hypotheses illustrates a theme of this chapter. Scientists usually discuss method only in terms of acceptance and rejection (themselves ambiguous concepts, as we shall see). In practice, they perform exploratory experiments and form as many conjectures and intentions for further experimentation as they do experiments aimed at contributing to a decision about truth. But this is kept under wraps, although if I am right it is relevant to some controversies about the meaning of the norms. So within the analysis there is a plea: be more upfront about your real intentions.

(2) *Record* as much as you can about the details of the experiment, well beyond the data that is analysed.

41 Bayesianism comes in two flavours. For subjective Bayesians probabilities are simply degrees of belief, and are thoroughly relative to the believer. Most philosophical Bayesians take this line. For objective Bayesians the source of the probabilities matters, and a lot of thinking goes into finding non-question-begging ways of determining them for probabilities of data and if need be prior probabilities of hypotheses. Most statistical Bayesians take this line.

It is worth distinguishing two main purposes among the many for this norm. First, it is inevitably frequent that an experiment supports a false hypothesis. The obvious reason is the statistical nature of much evidence. If something is true only of a small subclass of a population it is usually possible that a sample drawn from the population has disproportionately many from that subclass. And the experiment may have been conducted under conditions that are less general than is realized. When this is so literally replicating the experiment, re-doing it as near as possible in the original way, will often come up with a contrary result. Surprisingly many experiments do not replicate.[42] [Footnote.] But when an attempt at a literal replication fails it may be because of some detail of the original experiment. The source of experimental animals or materials may be different; the subjects may have been given different instructions. These things are rarely given in enough detail in published reports of experiments to enable a really literal replication. So the experimenters' notes may be invaluable, both for making a replication replicate and for diagnosing the differences between similar experiments with contrasting outcomes.

Of course a mechanically literal replication will involve the same procedures for analysing the data as the original one. There is a case, though, for literally reproducing the production of data while varying its analysis. One reason might be that the original experimenters' probability assignments were contestable, either because they had a partisan purely subjective element, or because they were based on a questionable model of the process generating the data. (There is more on such models in the next chapter.) This may be performed literally with the same statistical tests of the same data but different numbers. Or the intention may be nearer to making a case for an alternative

42 Saey (2015) gives references to a number of recent studies of the problem.

way of understanding the situation. This would be part of a larger undermining effort and could well have the limited ambition of showing just that an alternative hypothesis when used with alternative probabilities could fit this data. The emphasis then is on *possible* probabilities, which would be supported by an achievable model of the data production and which would make the data probable.[43] (More on how hypotheses suggest the probability of data in the next chapter (4).) These might play the role of reversal principles as in the previous chapter (2).

This prompts a very basic reflection. Doubt has closer links to possibility than probability. So when one is making a case for an alternative rather than trying to establish a claim it is appropriate to think about what is possible, or what might be probable, were this alternative the case, rather than what actually is probable given ones best estimate of the numbers. So wide-ranging scepticism, for example about the ambitions of science, is not best met with the observation that our conclusions are probably correct. We need a manageable version of the claim that we have good reason to believe they have some kind of necessary hold on contingent facts.

Further in this direction we get deliberately varied replications, varied in setup and materials as well as in analysis. The aim is to reveal the failure of the claim an earlier experiment seemed to support when applied in different circumstances or to a different population. These are sometimes called "conceptual" replications.[44] Then one is trying *not* to use the same design and materials although the concern is the same hypothesis, at least in verbal terms. Which aspects of the original experiments are deliberately varied

43 An example might be the recent use of quasi-significance-testing in particle physics, as with the large hadron collider. If the aim is to test extensions of the standard model then probabilities that presuppose it are appropriate, but if the aim is to explore alternatives to it they may be question-begging. Dawid (2015), Buckley (2016), Perovic (2016).

44 Stroebel and Strack (2014). I owe this distinction and this reference to Matthew Smithdeal.

will depend on the limitations aimed at. Again it will be important to replicate all the aspects that one is not deliberately changing as precisely as possible, and again this will often mean consulting with good records of the target experiment.

Experiments come in series, from small-scale exploratory experiments shaped around rather vague hypotheses to more specific experiments of more precise hypotheses, then branching to literal replications reproducing an original experiment as exactly as possible and to these "conceptual" replications. These may again be deliberate attempts to establish a hypothesis, though it may be a purely negative one, that an earlier result only holds under given circumstances. So here too we find a gradation in the definiteness that is aimed at, and again the attempts to provide convincing evidence are naturally associated with the probabilities according to the experimenters' best estimate of them, while the attempts at making a case for a possible alternative are naturally associated with the probabilities that might be found, for example according to plausibly different background hypotheses. But in this case the possibilities in question are those one gets by variation on more specific aspects of the model or expectations that were originally in force.

(3) S*elect* individuals for the experiment from a background population that you intend the hypothesis to apply to, and divide them into two groups. The experimental group has or has applied to it the attribute you are most interested in, and the control group does not. Try to make both have as great a diversity of factors that are not known to be relevant to the hypothesis in question as possible.

Part of the purpose is in a way obvious, and may seem blandly uncontroversial. Part is not, also, and discussing the un-obvious part prepares for some points about items later on the list. We are going to compare the two groups hoping that this will indicate something about the attribute, and we want to isolate it from other factors that might be relevant as much as we can. I have expressed this so that it does not require an experimental intervention (such as applying a treatment) and does not have to be directed at establishing causation. The reasons for expecting that overall similarity will help here are not quite as obvious, though.

It is easy to find obviously crazy inferences that would ignore this advice. Take an experimental group of people and write a magic symbol on their foreheads. You will find that all of this group eventually die, showing that the symbol is very powerful. Comparison with a control group whose foreheads are left untouched would prevent this. And if we do choose a control group, but of young healthy people, we will usually find that the treatment group has a lower life expectancy, also indicating that the symbol has a sinister power.

So why should comparing samples with and without the attribute provide evidence whether it is associated with some outcome? And why should this evidence be stronger if the two samples are otherwise similar? In fact, all similarities between the two groups are not equally relevant. Suppose that we are considering the relevance of age to a particular disease. Having groups that are matched for gender is likely to be relevant, because we know ways in which men and women age differently. However we do not know connections between age, the disease, and the first constant in someone's name.

There is no reason, either from our expectations or from the facts as they actually are, why groups that are similar in this respect should provide better evidence. Moreover some attributes of the subjects will be of greater relevance than others. Ancestry and previous medical history are likely to be more important than political affiliation and taste in music, themselves more important than how their names are spelled.[45]

These rankings of evidential relevance concern what might have an effect on the truth or falsity of the hypothesis. Experimenters can only use the best information they have on what will have such an effect. If this information is wrong, the evidence that it leads to is well conceived but misleading. Later generations may say "took to be a sign of" rather than "provided a reason why". This provides a constraint and a qualification on the assignments of probability that an experimenter uses. Probabilities that are unglued from comparisons of the ease with which events can as a matter of fact occur have a particular irrelevance to questions of how effective an epistemic strategy is. Probabilities give a different and finer grid than degrees of causal possibility, but different though they are, the two have to generally align.[46] (This applies both to finely grained quantitatively precise probabilities and to cruder approximate ones.)

(4) *Block* the application of treatments to subjects.[47] That is, apply the treatments to some but not all of the subjects, so that some get each treatment or possess some attribute and some do not. Arrange the applications and the rest of the process in a pre-

45 Influenza would be a candidate for the disease. Some other diseases would work less well because there may be a correlation with national origin of ancestors, and thus a greater correlation with names.

46 A minimal constraint is that a proposition never have a higher probability than one it is more possible than. It is easy to think of probability assignments that do not satisfy this constraint. The converse constraint is much less plausible.

47 Blocking is the standard term for this aspect of experimental design. To avoid confusion I usually refer to obstructing or inhibiting an unwanted causal process rather than blocking it.

planned structure, perhaps spatial, with an eye to what could go wrong with the experiment.

This is a refinement of the previous norm. It makes a manageable comparison between individuals with and without relevant attributes, by structuring so that comparisons can be made part of the analysis of the data. It is like the previous norm, making groups relevantly uniform, in that it responds to anticipated additional causes. But it is easier to apply when selecting individuals with or without particular attributes, or making them have them, is difficult.

There are many blocking designs, described in textbooks of these things. There have to be, because the relevant attributes will lend themselves better to some arrangements than others. Some designs would not be more effective in some situations, though, unless there were a hierarchy of relevant attributes. Some designs are ineffective, for some hypotheses in some circumstances, because they are unlikely to have any effect on the result.

(5) *Randomize the selection* of experimental subjects from the larger population so that neither you nor the causal gods can predict which objects have which attributes. This is mostly directed at unknown or unanticipated attributes, and gives some independence from prior expectations. Attributes peculiar to the sample will still have an influence but it will be distributed in the same way, with the same probabilities, as in the general population. There is some tension with the previous two rules, in that they encourage a deliberate distribution of attributes in the groups while this one encourages the groups to

have whatever distribution is found in the larger population. As a result, quirky features of the larger population, irrelevant to all or even acting against the hypothesis in question, may have an unwanted influence over the data. This may happen when the hypothesis concerns all individuals that could within easy possibility possess an attribute while the sample is drawn from the actual population with all its irrelevant features.

The tension is not mentioned in standard accounts of experimental method. Resolving it will inevitably mean taking degrees of anticipated causal relevance into account. One way might be to make a large selection from the population with a mind to the attributes that are best included and those that are best minimized, then drawing actual experimental groups from this. The first selection would be directed at known troublemakers and the second at invisible miscreants lurking in the phenomena.

A larger population from which the groups might be selected could fail to be ideal for an experiment directed at a hypothesis concerning that very population. The reason is that the hypothesis is meant to be a law, a regularity, true because of the workings of its objects. On the other hand, the actual population may be formed as a result of factors that are irrelevant to the hypothesis and is thus not typical of the full potential range of things of that kind. (For an extreme example, suppose that the hypothesis concerned the origins of life, and the only available samples, in fact perhaps the only ones to exist, were on the surface of the earth. This would be no problem for hypotheses concerning life on earth but might well bias the evidence for hypotheses about life more generally.)

Randomization is usually done with a table of random numbers, or a randomizing device such as a die or a roulette wheel. Their distinguishing characteristic is usually taken to be that the probabilities of the various assignments are equal, and independent of previous assignments. But this presupposes that events with equal or nearly equal probabilities will occur roughly equally often. Without this randomization will not serve its function. (So superstitious people who think that the numbers in their date of birth occur especially often should use non-standard randomizers, or use something like Lewis's principal principle together with some record of long-run occurrences to make their probability assignments match what they see occurring.)

(6) *Randomize the application* of the treatments so that neither you nor the causal gods can predict which get which treatment and which get none, and it would take a supra-causal god to make one experimental object get the treatment and another not. Again the intention is to safeguard against quirky unanticipated influences and focus on factors relevant to the hypothesis. There is a subtle difference with the previous randomization, though they are often not separated, which shows up when the hypothesis to be tested is explicitly causal. Then the aim is to make a situation where the suspected effect might be produced by few causes except the suspected one. Random application of a treatment inserts itself between the determination to apply the treatment or not and many or even all causes except that determination and its direct causes and effects, so that the treatment occurs on all causal chains leading to the phenomena produced in the experiment. That is the aim, at any rate. As a result we can discriminate between direct causation and the effects of common causes. (More on this in chapter 5.)

In terms of the themes of this chapter the immediate point is that a randomized application of a treatment usually produces a sample that is not found in nature and (patterns of) phenomena of a kind different from any exhibited by the population in non-experimental situations. That is a virtue. It makes it easier to isolate what is relevant to a particular hypothesis, which may itself help explain non-experimental events. Randomization also illustrates the theme about reversal conditions in evidence, which I say more about below.

Randomization can indirectly make an experiment fit a hypothesis *less* well. Suppose that we are investigating whether exposure to an environmental influence, which we can control, increases susceptibility to a rare disease enough to make it protect against it generally. So we need to form treatment and control groups, but since the disease is rare they are going to have to be large. And so, getting large enough groups given that the decrease in the occurrence of the disease is likely to be slight will require recruiting from a large population. Our first inclination, to use readily accessible subjects near the laboratory, would then require more effort and expense, so we expand the pool by also working with people from the other side of the country. But what we do not know is that the genetic composition on that far side is not typical of our country, so that the environmental influence will have a lesser effect there. Not knowing this, we select from the larger pool and find that the influence makes very little difference. But if we had stuck with the first inclination and simply used as many nearby people as we could readily recruit we would have seen that protecting people against the influence would be

worth the cost. The now familiar point is that you often do not know the full nature of the evidence you have.[48]

(On a more subtle version of the point the probability distributions for the population of interest are different relative to different intended applications. Disparity of causal origin is a root of this. A topic to return to in the next chapter.)

(7) Use a *placeb*o whenever there is a possibility that the fact of being treated rather than the treatment itself is among the factors responsible for the data. (Unless that is what the experiment aims to show, of course.) This also can often create a situation that is not common outside experiment, since in other contexts the effects of expectation and treatment are often too tangled to be separated.

But analogues of placebo-use do sometimes occur in extra-scientific inquiry. If you want reassurance that the flight semi-audible announcement was of the flight you have been waiting for all day you ask the opinion of people not waiting for that one. If you want an objective opinion about whether you are overreacting to a situation you present it to a friend who you trust in a way that does not reveal whether you are involved in it. (The friend often sees through this.) Note however that in doing this you are choosing to ignore possible sources of information. People in the airport not waiting for that flight, perhaps not even travelling, may be tuning out the annoying announcements. The trusted person will not be able to factor in your particular reactions.

48 This is related to points about randomization made in Grossman and Mackenzie (2005), which argues that the advantages of randomization often do not outweigh those of other statistically desirable qualities.

Such problems apply less with experiments in scienc, because there is usually only one source of information. But they are still possible. They will occur for example when human subjects are given instructions that need some interpretation. The experiment might concern how well people pick up hints about an object. Two hypotheses are being compared. One asserts that we most readily interpret hints about familiar objects, and the other incidents that we do best with animals. In the experiment one person gives cues about the kind of object they have in mind, and a second person produces a candidate kind to match it. We have the option of doing this also with unmatchable descriptions, which are neither of familiar objects nor of animals but of remote and unfamiliar things. These serve as a kind of placebo, preventing the second person from eliminating possible answers in terms of the purpose of the experiment. But if we use this placebo we are excluding a significant source of people's normal interpretations, namely their understanding of the reasons for an utterance. This creates a distance between the experimental results and outside the laboratory they are meant to illuminate. This relativity to context frequently qualifies the planning and results of experiments, where we have to choose between aiming at underlying processes ("deep causes" in the terminology of chapter 5) and aiming at reliable patterns outside the lab. But it is also telling that this is an issue about the causal origins of the data. Placebos shield against one source of data, so we have to decide whether this source is relevant. Are we interested in discerning some particular factor shaping them, or in getting a comprehensive picture of the variety of their causes?

(9) *Blindness* applied to experimental subjects is helpful when there is any chance that the expectations of subjects, experimenters, or others involved in the experiment, can influence its outcome.

This is single blindness, where human subjects do not know whether they are receiving the treatment or not, but the experimenters may know, and serves much the same function as administering a placebo and has similar rationale and limitations. Subjects may be kept in ignorance in more general ways. For example they may be given a deliberately misleading interpretation of the purposes of the experiment. This is common in psychology. Again there are specific factors that are excluded, and this creates an unusual situation directed at a particular theoretical purpose and sometimes interfering with others.

There is a slight tension between single blindness and randomization. Blindness might ask for subjects from a group, some of whom might understand the purpose of the experiment in inadmissible ways (suppose that we are studying whether psychology professors are less susceptible to subliminal influences). But if we select randomly from this group we will get these unwanted subjects. The obvious remedy is to test for the wrong kind of subject first and then to randomize.

(10) Sometimes *double or triple blindness* helps the purpose of an experiment. In double blindness procedures many of the people carrying out the experiment do not know which treatment has been given to whom, and in triple blindness some of the functions of data collection and analysis are separated from the final conclusion-drawing. For example the

classification of subjects reactions in a psychology experiment may be handed over to assistance who do not know the true purpose of the experiment. This protects against self-serving interpretation by the experimenter with a favoured hypothesis to support. Again the situation produced is unnatural, though in this case the fact that it is an experiment testing a particular hypothesis is part of what is unnatural about it. And again the result is a delicate mixture of advantages and disadvantages, with a balance that depends on the purpose of the experiment.

reversal principles again

In a previous chapter, (2), I introduced the idea of a reversal principle, which derives one conditional "if data then hypothesis" from another "if hypothesis then data" given certain conditions. I suggested that experiments provide conditions where these are true. Do they? Adherence to these rules very often does.

The general idea amounts to the Sherlock Holmes principle.[49] If something can have many causes then restricting their number or range or influence increases the likely causal force of those that remain. To put it explicitly but still in a preliminary way, suppose that e can be the result only of a finite range of identifiable causes C_1, C_2,..., C_n = H,, where H is the hypothesis we are testing. Suppose we have eliminated C_1,…, C_{n-1}. Then the only remaining possible cause is C_n. So when e is caused to appear in an

49 Often quoted, from "The Adventure of the Beryl Coronet" where Holmes says "It is an old maxim of mine that when you have excluded the impossible, whatever remains, however improbable, must be the truth." Note the slide between possibility and probability. Note also that especially when formulated purely probabilistically this involves a balance between the force of the exclusion – how improbable the excluded factor becomes – and the initial improbability of the conclusion.

experimental context that has prevented $C_1,\ldots, C_{n-1}$, then C_n is the cause and thus H is true. But this is the non-comparative reversal condition

$X \rightarrow (((C_1 \text{ v } \ldots \text{ v } C_n) \rightarrow e) \rightarrow (e \rightarrow C_n))$ where X is the background conditions and the situation set up by the eXperiment.

That is the general strategy. But it is much too rough as stated. To formulate it a little more carefully note first that the C_i describe causes rather than arbitrary conditions. It is more plausible that causes are finite in number (proximal causes, anyway). And it is more often true that an experimental situation can eliminate some of them. Still, in real experiments one diminishes but not eliminates the influence of unwanted causes (another reason that statistical apparatus is inevitable) and even drastically reducing all but one is rarely a possibility. But we can stick weights on the C_i, without changing anything essential. The weights can represent probabilities, relative causal influence, or normalized products of probabilities and influences. Then the presence of X can change the weights, proportionately increasing that of C_n while reducing the others.

So let $w_o^{oi},\ldots,w_n^{oi}$ be the weights that the causal factors C1,…,Cn have when they produce Hi (i=1,2), or its effect that is the proximal cause of e outside an experimental context, and $w_o^{wi},\ldots,w_n^{wi}$ those that they have *w*ithin it. If the setup is at all competent this X has the consequence that

$X \rightarrow ((w_o^{o1} C_1 \,\&\ldots\&\, w_n^{o1} C_n) \rightarrow e_1)$ and $X \rightarrow ((w_o^{o2} C_1 \,\&\ldots\&\, w_n^{o2} C_n) \rightarrow e_2)$.

where $\sum_i w_i^{w1} = 1 = \sum_i w_i^{w2}$ and for $j \neq n$ $w_j^{w1} < w_n^{w1}$, $w_j^{w1} < w_n^{o1}$ and $w_n^{w1} > w_j^{o1}$. (That simply requires that X reduce all the unintended causes to the benefit of C_n.)

Then the same reasoning as used in chapter 2 applies. (Remember that C_n is H_n or its consequence that is proximate to e_n.) If the result of the experiments is e_1 then the incompatible e_2 did not occur, so if one of H_1 or H_2 is true then it is H_1. Or put more formally

$$X \rightarrow ((w_o^{o1} C_1 \,\&\ldots\&\, w_n^{o1} C_n) \rightarrow e_1) \;\rightarrow\; (e_1 \rightarrow C_n)$$

This is the reversal principle for experiments described in terms of weighted causes. The formal versions have the advantage making explicit how the reversal relies on the presence and properties of the weights. If they do not have the required properties then the reversal will not work. The point that C_n is the proximal cause on the route from C_n to e_1 is particularly important. Experiments do not always tell us why the immediate causes have the effects that they do. It is also significant that this reversal allows for the less than total reduction of all but one cause, so that the result is subject to the inconstancy of the others. It shows us not that something is always the case but that it is often so, true with a given probability.[50]

There is a delicate balance here. On the one hand a typical experiment gives information about a proximate cause under the experimental conditions only. On the other hand the tunability of experiment allows us, with ingenuity and luck, to focus on deeper reasons why the apparently proximal causes have their effects. This often means that the results of an experiment apply best to a particular domain under particular conditions.

50 Again there are issues of how well an experimenter can know what the force of the results are, and in particular whether the inequalities among the causal probabilities (propensities) hold. But it is striking that this context modal considerations (which situation the parts more from actuality) and probabilistic ones work together (though for a rather particular interpretation of probability, as measuring causal tendency).

cheating and the force of evidence

What is the price for breaking the rules? Consider "cherry picking" — forming your beliefs or what you tell others only on what results from favourable studies[51] — or planning in advance to cease a trial when a parameter reaches an intended number. These may result in true beliefs but the "evidence" they are based on will not be objective (sensitive, robust) evidence , and what it leads to will not be knowledge.

Putting the point this way suggests another class of cases. Suppose that the basis for the evidence is not what we think. What we took to be a placebo is actually easily recognized as such. (I gather this is not so rarely the case.[52]) An over-zealous assistant has culled the animals who did not seem to be doing well. A computer file has got corrupted. None of these things suggest that the researchers are dishonest or that they are reasoning badly. But they do suggest that the claim to knowledge is mistaken; the conclusion is well justified and perhaps true of many cases, but not known. Conclusions formed like to this in similar cases will often be false, however sensible or rational the person was in using them.

But which violations of which norms present the greatest threats? The relativity of experiment-backed conclusions to intended domains suggests a classification. Contrast two extremes. At one the evidence is for a hypothesis very generally, in a wide range of populations in many circumstances. (An example might be an experiment in molecular biology to determine an aspect of the gene-expression mechanisms that fits daughter

51 Tests sponsored by drug companies are often accused of this sin.

52 As a student I was a volunteer subject in a psychology experiment where it was pretty clear that the hypothesis in question was unrelated to what we were being told.

cells to synthesize the proteins of their parents.) In such cases further experiments testing the hypothesis under variant conditions will be relevant, so very wide record-keeping makes sense, to facilitate "conceptual" replication. Some of these experiments may well draw from different sub-populations. And in general provision against the full range of influences, including unknown ones, will help wide export. Randomized selection of experimental groups tends towards making them homogeneous with the larger population, and so it is a standard such provision. It is high on the list of desirable features when context-independence is the aim.

At the other end of the spectrum, extreme context-dependence is approached the more the sample is taken from a very particular sub-population influenced by factors peculiar to that sub-population or even that sample. (An example might be an experiment to discover the breaking strength of a particular suspect girder manufactured by a process that has many variable and even random elements.) In such cases the focus is on populations much like those in the experimental groups, and subject to influences much like those in the experiment. More precise replication is then appropriate, and information about the details of the experiment is worth preserving. Inasmuch as the description of the population and the experiment is accurate, we will have more information about the specific factors that will have been taken account of. Put differently, there is potentially a fairly complete list of what is allowed to vary from case to case and what the experiment tries to keep constant. Blocking techniques tend to keep this control. Accurate background information about which situations are causally similar to those in the experiment is obviously useful here, when we can have it. When

the parochial domain is human beings placebo and blindness measures can be relevant to keep unwanted influences from beliefs and expectations at bay.[53]

Between the extremes the balance is, well, between. It can be summed up with a table.

wide export	**mixed**	**narrow export**
conceptual replication	replicate towards application	exact replication
\\\\randomized selection	\\\\\\\\\\\\/\/\/\///////////	blocking////
.....placebo, blindness.....		

There is a connection here with the issues about error versus ignorance discussed above. But what emerges is more subtle, and suggests some finer tuning of the distinction. By choosing which influences to shield against against and which to leave functioning, to the extent that we can, you decide the purposes of your hypothesis acceptance. If you are fishing for promising beliefs in a certain area you set up experiments so that they are particularly sensitive to the relevant influences, and shield against those that are more important elsewhere. So you can choose which kinds of ignorance you want most to avoid. The issue is broader than simply what kinds of hypothesis you accept. Accumulating evidence for interesting possibilities is equally valuable. And using reversal principles, even without the probabilistic ideas that become increasingly important in later chapters, need not lead straightforwardly to acceptance or rejection. For what you get from such a principle is that an experimental setup is sensitive to a certain kind of fact. You can decide how sensitive you want it to be, in terms of the alternatives you compare with it and in terms of the fine-tuning of the conditional that you use. In all

53 Sometimes these influences are what we want to understand. An experiment may suggest that a folk remedy has no strictly medical value, while at the same time its aura gives it effectiveness for many people.

these ways choosing how you distribute emphasis on the various desiderata of experiment allows you to influence the kinds of hypothesis that survive an experiment.

This chapter has focused almost exclusively on scientific experimentation. In everyday pre-scientific life there is also a tension between knowledge-directed interventions that aim at a specific situation and those meant to tell us something more general. Common sense is notoriously unreliable when applied to unfamiliar situations. When we learn from acting on things our knowledge usually applies best to those particular things, the people and physical objects that we are in touch with. Its failures to assess causation are particularly salient. Though I shall not do more than state the claim, it is very plausible that pre-scientific acting to know is most like experimentation in a tightly controlled context on a specific small population. More like the girder case than like gene expression.

chapter 4: robust tests

probability theory meets modal semantics

Experiment is where data meets doctrine. They have to reach a compromise, but it may take a long conversation. Procedures for testing hypotheses against the results of experiment are at the heart of most experimental research. But these procedures are controversial and factional. They are contested in the "statistics wars" between adherents of different conceptions of probability and associated schools of hypothesis testing. Yet the practice of most applied statisticians is rather eclectic, drawing on ideas coming, sometimes inconsistently, from the warring schools. This suggests that lurking behind the practices there are underlying unstated principles that are the same whatever line one takes on the same time more contentious issues. This chapter is a step towards formulating them.

The chapter is shaped by an attitude to the reevaluation of hypotheses in the light of evidence, from an early judgement whether they are worth further investigation to a late decision to take them as established. Talking of confirmation insinuates without argument the idea that evidence works the same way at all stages of this process, something that ought to be seen more critically. Still, the assurances that we want in the evolution of an idea from conjecture to orthodoxy concern various aspects of the conditions under which a proposal could in fact be true, given the actual often unknown facts. We want evidence to push us in the direction of greater truth, nearness to knowledge. So theories and inferences can be less secure than we intelligently and honestly suppose. A theory can be

false, unbeknownst to us, and an inference can be risky. In particular we want to link probabilistic accounts of hypothesis testing to facts about the conditions under which the hypotheses could or would more or less readily be true. Probability meets possibility.

The epistemological theme of this chapter is the stages between curiosity and acceptance, and the different tunings of evidence that they require. (Including the tunings of probability.) These are also themes for the philosophy of science, though with an emphasis on the procedures used in contemporary science.

Metamorphoses, egg to larva

It begins as curiosity and hatches as conjecture, sometimes a fairly definite conjecture and sometimes not more than a feeling that something of some kind is happening, perhaps that some pattern is not an accident. So you begin to investigate. (I shall leave it open until chapter 6 whether "you" is singular or plural.)

You are going to develop the idea experimentally, so the first step is to construct experiments indicating whether there is anything going on and if possible what kind of a thing. (Whether a condition might be a result of a bacterial infection; whether dark matter involves new fundamental particles.) The results may suggest a more focused conjecture and your next task is to get this into shape for whatever explanatory or practical task is intended. This may involve uncovering the causal processes behind the effect, relations with other better established processes and principles, and more careful delineation of what is and isn’t being suggested and under which conditions. All of these

may involve factual discoveries and modelling of subsidiary mechanisms. Somewhere along the line comparisons with variant or competing formulations are likely to be helpful. And then in a final stage the overall strength of the evidence must be assessed and the surviving doctrine must be woven into the fabric of established belief. The process can take weeks, months, years, lifetimes. The very final mature version can be short-lived, lasting just long enough to engender another cycle of ideas.

The process begins with a rough idea that asks to be either developed or forgotten.[54] Take the question to be "is anything going on here". The possibility to be ruled out is that some apparent pattern is entirely accidental or illusory. (These are different, of course; some accidents are very solid.) If this is not so then it might be supported by experiment under controlled conditions, including the production of examples that would not naturally have occurred. The data of such an experiment must be analysed to compare them with natural variability or chance. The often-maligned significance test has honest employment here. You first construct a statistical model — details below — which gives probabilities that the appearances are indeed due to chance. Then you see how probable or improbable the experimental data would be if this particular chance mechanism had produced them. You judge this in terms of a fixed level of improbability, a p-value below which you will not take the idea seriously.[55] This much is familiar, but it leads to a chain of further points.

---

54 "Idea" is deliberately vague. It could be a suspicion, a conjecture, a working hypothesis, a best available explanation, a solid part of established doctrine,… . Theory-like vehicles that are not final and self-contained like the older concept of a theory are a theme of recent philosophy of science. Cartwright (1994, 1999), Morgan and Morrison (1999), Wilson (2018), and even in a way Van Fraassen (1980).

55 Garthwaite, Jolliffe and Jones (2002, section 4.3). Most of the my statistical references will be to this work. It is a somewhat sophisticated and compressed exposition; similar topics are covered in Lehmann and Romano (2005), Wasserman (2004), and most accessibly in Bulmer (1979). I am supposing a situation where an upper limit is appropriate. Different situations may call for a lower limit or limits at both extremes.

Many different factors can be called "accidental" (or "random", or "chance"). Even when one concludes that there is some phenomenon needing an explanation the data will almost always have some degree of irrelevant variation — the line almost never passes through all the points precisely — which is usually due to the effects of other causes and to errors in measurement and observation. Sometimes it is due to inherent randomness in the correctly hypothesized factors. So the aim is not so much to eliminate unexplained variation as much as to diagnose it correctly. Statistical models play a role here also.

Even when one has distinguished between variation stemming from the hypothesized factors and from other sources and made a case that something real is going on, or even what kind of a something it may well be, an experiment like this is not going to reveal much about the details. There will invariably be an array of alternatives to the "there's nothing here" option. (Although a conclusion that there is something to be understood better, for example that a fertilizer improves crop yields or Fisher's example of the woman who claims be able to tell whether milk or tea was poured first, can be surprising and look rather as if it was an explanatory theory.[56]) So this first stage, and this kind of testing, can only be the beginning of an investigation. Most of the well-known abuses of significance testing stem from taking it as if it were appropriate further along in the process.[57]

Assuming that the investigation will be continued, the criteria for survival, "passing" the test in this particular way, can be fairly permissive. p, the confidence level, does not have

56 Fisher (1935). Note that Fisher has her fail the test, so that there is no (causal) phenomenon to investigate.

57 Cohen (1994), Gigerenzer (2004), Wasserstein and Lazard (2016).

to be extremely low. It can however be varied in accordance with the financial or other cost of proceeding with an investigation.

Keeping these points in mind, tests like this are not completely different from some pre-scientific procedures. We often consider events and conclude that they cannot be a coincidence. All your five best friends have unusual quirky and sometimes incompatible senses of humour yet there is a joke that appeals to all of them. If you think that just about impossible that an ordinary joke happens to hit all their buttons then you will wonder whether this is an especially powerful joke, of almost universal appeal. This does not mean you will believe the conjecture, but that you will try to remember it and try it on others. Otherwise you will just put it down to chance.

A different application of a similar idea outside of science arises with the modern realization that many things that result from chance cannot be attributed to human or other agency. They just happen, and there is no point looking for deeper reasons or motives. So when tragedy strikes good people rather than villains it is not because of some mysterious choice of the gods but simply because the causes of tragedy have nothing to do with the causes of personal value. The idea that profound powers have it in for the affected person is a nonstarter, since what happens is at least as likely on the milder assumption of randomness.

At this point there is a connection between the everyday and the abstract ends of the spectrum, more specifically with a frequently cited, though puzzling, criterion for hypothesis tests. The controversial “likelihood principle” requires that hypotheses be

judged just on the actual data, not in comparison with what might have occurred.[58] But consider how you would react if you meet a friend, who always dresses very formally because it is required for his work, in the middle of the day downtown in shorts and a T-shirt. You wonder whether he has changed employment or lost his job. (You do not conclude that either has happened, but rather that you should ask him or otherwise find out.) You have this thought because of what you do *not* observe, more formal attire. The basis is the absence of the most likely consequence of what you would have expected.

The moral to draw is that the likelihood principle is not a good guide when the purpose of a test is to clear an idea for further investigation. This should not be very surprising, since the upshot of such a test concerns an indefinite might-be rather than a definite claim on fact.

The thumbs-up that a significance test can give to an idea is pointless unless the idea is at least possible enough to be worth investigating. The decision is made in terms of probability, from probability distributions based on a comparison of what the idea claims and what is expected if the results are due to chance. The probabilistic aspect of both may well need to be made more explicit enough for a numerical comparison. This is the task of a statistical model, discussed below. The limited aims of such a model, for this investigation-prompting purpose, require at a minimum that when the suggestion is more probable than chance it is at least as possible: if only one of the suggestion or chance production is true then the suggestion is a better bet. If this is not the case then the assignment of probabilities to events is unconnected with what it would take for them to occur, and the whole procedure is hollow.

58 Birnbaum (1969, 1972, Grosman (2013), Gandenburger (2014)

pupa to larva

Once the germ of an idea has acquired enough support that we are ready to investigate it further, it needs to be brought into shape for more conclusive evaluation. So it must be formulated more precisely and the more precise versions must be compared. A central aim of precision here is to capture more details of the data, although many details will not be available until experiments to choose between variants are designed and performed. So we can expect that in general there will be a number of variant versions, and experimental ingenuity will be needed to compare them. Comparison is the main aim, and experiments are directed at this. Comparison and likelihood are good companions. The likelihood of a datum relative to a hypothesis is its probability given or conditional on that hypothesis. (It is different from conditional probability in that the likelihoods of a set of mutually inconsistent and exhaustive outcomes, given a hypothesis, need not sum to 1.) Comparing the likelihoods of the same data on two different hypotheses — in particular now two different variants on one guiding idea — tells us a lot about how well each hypothesis handles the data in comparison with the other. The comparison is most commonly summed up by the ratio of the two.[59]

There is serious work to do before this is possible, though. Much of the work involves constructing two kinds of models. The first kind, mediating models, connect theoretical ideas with observable phenomena. The hypothesis might concern the effects of a drug on an infection and it might need assumptions about how quickly the drug is absorbed if taken in a particular form, how rapidly it diffuses, and how long it persists. If an

59 Hacking (1965), Pawitan (2013).

experiment is performed using such a model and it gives no useful results, the model can be tweaked and the experiment re-done while still testing the same theory.[60] (In a different application, or as inspiration for a different theory-pupa, a different model might be appropriate.)

A mediating model describes the connection between the processes that a hypothesis supposes and some class of phenomena. But it typically does not address the actual distribution of outcomes and the extent to which they conform to the postulated influences. For this we need a statistical model. It can consist simply in a set of probability distributions, traditionally given in terms of a range of one parameter in a standard formula for a family of distributions. Then a particular narrow range of values for the parameter is associated with the specific hypothesis. If it is to be compared to an alternative then that alternative will have its own range of parameter values. These two ranges are important. They provide the probabilities for the two likelihoods. (I discuss this further below.)

Researchers usually construct their own mediating models. But statistical models are often the work of statisticians not centrally involved in an experiment. They discuss with the experimenters the kind of process involved and the differences between the two hypotheses and then produce the model, which is often just two ranges of a parameter in a common distribution. The doctrine behind such model-making consists largely of statistical folklore; there is very little in the way of general explicit theory.[61] (If it existed,

60 Morgan and Morrison (1999). The essential point is that such models are directed at capturing the data in a particular application, and at making this data manageable instead of stating in anything like completeness the causes of the data. That is why variant models are not incompatible; they may do the same job in different ways.

61 Cox (1990), McCullagh (2002), Mayo (2018) section 4.8 discusses the simplifying ("false") aspect of statistical models.

it would amount to a kind of quantitative scientific metaphysics. So it is not surprising that there is little to be had.)

Armed with models of these two kinds, we can test hypotheses comparatively. There are competing accounts of how to do this. They differ mostly on the source of the probabilities and on whether probabilities for the data are needed, as they are on Bayesian approaches. In standard non-Bayesian methodology "prior" probabilities of data are not necessary, as likelihood ratios do not need them. And on these traditional approaches the probabilities actually used in the analysis of an experiment are derived from the hypotheses themselves. That is, they are intrinsic to the statistical hypotheses that the model associates with the more substantive hypotheses of association or causation. The analysis of the results of the experiment is directed at evaluating which distribution — so typically which parameter values — best fits the data. The heart of this is often a likelihood ratio where the likelihood of the data given each hypothesis is given separately, since it is actually the statistical targets, such as the ranges of parameter values as supplied by the statistical model, that are compared[62]. (One effect is to hide from experimenters that there is any issue about the origins of their probabilities, since they seem intrinsic to the hypotheses and the transmutation of the hypotheses really of interest into claims about parameters and distributions is handed over to the statisticians who simply appeal to their collective wisdom.)

A frequent form for this stage of development of a scientific ambition is thus the comparison of variants of a common idea with assistance from improvised models of both mediating and statistical kinds. Here too there are precedents in pre-scientific life. And

62Garthwaite, Jolliffe, and Jones (2002) 4.6 & 4.7, Lehmann and Romano (2005) ch. 1.

the anticipation of comparative confirmation and the anticipation of modelling are also found. Suppose that you have a suspicion that someone is hiding something in their descriptions of their past. That is a general pupa of an idea, but with a few vague indicators on their part you can take it seriously enough to move to the next stage. You are then comparing the possibility that they are deceptive about what they are telling you about past accomplishments and qualifications against the alternative that they are a modest person who becomes awkward and indirect when they might admit to what someone else might boast of. Since you cannot summon an array of honest testifiers about all stages of this person's life, you choose a particular domain, educational achievements. You focus on the rather sketchy description of education in their CV, and you prepare some pointed questions about one of them, on which you have expert knowledge. You are proceeding on the assumption that if you begin with general bland questions and then advance to sharp precise ones the person will begin with a confident cluster of replies and then retreat to confusion and obvious obfuscation if the first hypothesis is correct, and become more and more focused if the second.

There are similarities to the three main elements of the pupal stage of science. There is the progress from general suspicion to more detailed hypothesis, in this case from a vague imputation of evasion to a hypothesis about qualifications. There is the exploration of a contrasting pair of ways of working out the idea in more detail, as a way of getting a confirmation or refutation. And there is the temporary adoption of a possible mechanism in a possible application in order to make an experiment possible. Last of all, there is creating a situation where particular events may indicate which of these two hypotheses

is more likely. All three suggest a profitable strategy which has widespread applications when we want to set up an information-rich situation.

interruption: statistical models and the sources of probabilities

I interrupt the sequence of metamorphoses for some remarks about statistical models and probability. The concept of probability plays a central role in many, probably most, hypothesis tests, especially in the form of likelihood, and especially especially in the form of comparative likelihood. (The roles of prior probabilities of theories and of items of evidence are more contested.) It is hard to see how we could function without it. But where do the probabilities used in hypothesis testing come from?

Most frequently from statistical models, so usually a class of distributions and a range of parameters. Sometimes the model is nonparametric, when the distribution is specifically fitted to the situation. The choice of distribution and parameter tends to be made on the basis of a rough sense of what is going on plus statistics of existing populations. Again agricultural and medical hypotheses are an example. The distribution of the quantity of interest is usually Normal, by extrapolation from the untreated population, and since there are many independent sources of variation from one case to another a Normal distribution makes sense, appealing to the central limit theorem. So in the simplest cases one associates the null hypothesis "nothing special going on here" with a Normal (or polynomial) distribution with mean and variance as in the population sample, and one associates the hypothesis that the treatment has an effect with a similar distribution but

a greater mean. Somewhat more probing but still rather schematic considerations will lead to other families of distribution.

Once the model is determined, or at any rate decided on, there are rich enough probabilities to define the likelihoods. So from that point on the experimenter does not have to think where they come from, and in particular can avoid grappling with thorny issues about the nature of probability. These issues are themselves avoided if we take a particular attitude to the models, especially if they are fleshed out with a description of the causal processes. I shall call this the Neyman attitude.[63]

To describe the attitude take the model to consist of an infinite population of individuals with a very limited number of attributes, where the proportions with each attribute are as the hypothesis or the statistics of the actual population would suggest. This is clearly a model, a simplified abstract structure, rather than a description of the physical facts. The infinite cardinalities are no problem since we are dealing with an abstract, essentially mathematical, structure. And while proportions of infinite sets are problematic, the limits of many sequences will be well-defined. Worries about sequences that do not converge to limits and about selections and rearrangements from sequences that do not behave in comfortable ways are irrelevant. The attributes that define the sequences are chosen so that they correspond to real properties of the physical system being modelled as related by the causal features of that system. For example we can model a large number of throws of a fair die with a model in which it is thrown an infinite number of times. The proportion of 2s to evens will approach 1/3. A straightforward way of making the model

63 Neyman (1957), (1990). See also Strevens (2011), I am sure this is not the application Strevens had in mind.

mimic and extend the probabilistic aspects of a physical system would be to build it around a random number generator, so that it would in effect incorporate a collection of versions of each probability-governed sequence.[64] The price, however, is that the model must itself be abstract and contain no more structure than is specified, and at the same time correspond to the physical workings of the system.

This requirement of correspondence to the physical system constrains the probabilities. It rules out probabilities that are simply degrees of belief. For these need have no connection with the actual workings of a physical system. And if they are intended to represent its workings but do so inaccurately, then to the extent that they differ from the actual tendencies of the system they are defective, though the experimenter using them may not know this. Moreover if we link them to mediating models saying how the more general features of a hypothesis connect with a particular class of phenomena the obstacles to such purely methodological probabilities become greater. The opinions generating the probabilities would have to be manufactured for this particular purpose. I cannot see what the basis for the manufacture would be except for some version of of the attitude that probability measures how often an event would occur if the process generating it were indefinitely repeated. (The wording is deliberately vague.[65])

We are often or typically comparing two hypotheses. Then the likelihoods of each are given by each itself, with a lot of help from the statistical model. The aim is to reveal which is nearer to the truth. Of course, but nearness to the truth can be understood in

64 The result would be like a model for a modal logic, with a set of first order models for each satisfiable sentence bundled together by means of an accessibility relation. To fit standard accounts of probability accessibility would be an equivalence relation and the result would be rather S5-like, but I wonder whether a more general approach, with milder conditions on accessibility, would be profitable.

65 For reasons to keep things open see Hájek (2009). One reason for being notably cagey about probability, besides the difficulty of any other course, is to balance the lack of consensus about possibility.

various ways. The direct link to it is through these probabilities. But there is a complication. The link will be weaker the less accurate the probabilities are, and since the hypotheses are typically incompatible we know that at least one of them is wrong, and unless the probabilities they supply are different there will be no test. So there is some suspicion over them. Since we are using both distributions, we are condemned to using at least one faulty one. This will push us in the direction of smoothness of fit with the data rather than correctness of explanation, which becomes harder to justify as a means to truth the more remote the link between the hypothesized processes and the facts becomes. It is pretty plausible for a low level data summary and much less plausible for a complex causal explanation. A minimal requirement is that greater probability correspond to greater possibility, in the sense of nearness to actuality. Unless this is satisfied there is little connection between probability and what is more and less prone to occur, and the procedure is hollow. [66]

A natural form for this minimal requirement to take is as a comparison of likelihood tests. Let us say that one test comparing H_1 to H_2 is better than another when it gives H_1 a higher probability than H_2 when H_1 is true if (H_1 or H_2 and not (H_1 and H_2)) is. This will automatically make a test that prefers a true hypothesis on the basis of its probability better than one that rejects it in favour of a false one. The requirement will also fault many tests that are susceptible to overfitting, giving disproportionate weight to tiny details in the data that can easily be produced by random fluctuations. An overfitting test might for example prefer a convoluted squiggle over a straight line passing very near to a large number of data points but not going precisely through many of them. If the underlying causal principles are linear then we will want the test to prefer a hypothesis

66 Strevens (2013) gives a careful and technical formulation of one such condition.

giving a straight line over one giving such a squiggle.[67] This will be so not only when the first hypothesis is true and the second false but also and more generally when the nearest world in which the first holds is nearer than the nearest for the second.

This requirement is obviously externalist, in the sense that it depends on what is actually true and how possible something could be, rather than what a reasonable and adequately informed person might think and do. We need criteria of both kinds and of course we try to bring them together. Rules such as David Lewis's principal principle are designed to take us in this direction. There could more generally be a relation of betterness between probability assignments which prioritized their congruence with degrees of possibility. But the immediate focus is preference between theories, where the concern with truth is paramount, and probability is used for a wide variety of purposes. The topic will return when we discuss the severity of tests later in this chapter. (I am using this section to separate the discussion of appropriate probabilities from that of theory-acceptance and that of severity partly because they should be clearly separated, and partly because the content of this section is more relevant to the previous one than to the next.)

larva to adult

When a hypothesis has survived a series of tests against variants and rivals, there is still a question of whether it should become part of accepted scientific doctrine. The tests may not be enough to establish this, for a number of reasons. For all its success with the data, the hypothesis may not have been tested against some relevant rivals, perhaps because no one has been able to devise suitable experiments. (Experiments have failed

67 **Forster and Sober (1994)**

to find real processes and entities because they were mistaken about how to detect them.[68]) It may not fit nicely with long-established and well confirmed theories. The tests in the series although individually successful may undermine one another or reveal one another to have very little force. It may simply be implausible. (Basic implausibility obviously overlaps with tension with other longer-establiised theories and beliefs.) And when designing statistical models the data from existing populations is often not enough; for example we need to anticipate how large an effect a novel factor may produce if it exists.

The most interesting aspect to a comprehensive theory may rest on the way it unifies more detailed experimentally supported suggestions although it is not easy to test it by experiment directly. The quark model of hadrons is an example, and outside science the picture of the mind as integrating beliefs, desires, and feelings is central to our social attitudes but is hard to test as a whole. So a complicated and delicate basis for accepting or rejecting a proposed addition to received wisdom is how it combines with other more constrained results as parts of a comprehensive account of a range of phenomena. Having mentioned this I am going to say very little more about it.

All of these are reasons for wanting to assess one piece of theory from the point of view of another. They require the force of their credentials to be compared. Much of the time this is a purely intellectual process, but sometimes properly experimental considerations are relevant. They apply when the results of an experiment need a lot of interpretation to provide evidence about a theory. The detection of gravitational waves is an example. The

68 Galison (1987) gives the example of Maxwell's attempts to see whether the carriers of electrical charge have inertia, which failed because he misunderstood the range to look in.

bare data could be understood many ways in the light of various actual and possible physical theories, but assuming that general relativity is probably correct, the data give a high probability to the predicted distortions of space-time by the behaviour of massive objects. The same could be said of routine interpretations of MRI output, and for that matter of a great number of imaging studies where a sophisticated diagnosis is required to get the image from the output of the apparatus.

An additional and perhaps flimsier reason is to give information about whether an interesting idea (a pupa) is worth testing at all. The probability of the idea given a well-established theory can be helpful. (Except when the idea concerns how the established theory could be fundamentally wrong!)

The aspect of these theory-presupposing tests relevant to this section is their essential use of prior probabilities. We need to factor in what are more and less likely explanations of the data, and this has to build on estimates of how plausible these explanations are before the experiment, so prior probabilities. This does not have to be full-scale classic Bayesianism, though that would do the job. The prior probabilities can come from some other source than degrees of one person's belief, and in fact in much of statistics "objective" Bayesian methods are not degrees of anyone's belief. They do not have to be numerically precise, for many purposes. And for turning experimental results into evidence we do not need more than ratios of prior probabilities ("Bayes factors"). Then we can evaluate the ratio of two hypotheses' probability given common evidence as

$$(H_1|e) \,/\, P(H_2|e) = P(e|H_1)P(H_1) \,/\, P(e|H_2)P(H_2)$$

which is the likelihood ratio weighted by prior probabilities. This is usually enough to say which one is better supported by this particular experiment, and if you have a lot of faith in the numbers by how much. Besides likelihoods it needs prior probabilities of the hypotheses, but not of the evidence. (Which is just as well as this is notoriously problematic.[69]

The evaluative strategy at this final stage can be taken as the culmination of a series where each stage provides material for the next. It starts with a single hypothesis compared to the results of chance, which then passes through tests against an alternative in terms of comparative likelihood, and now to evaluation in comparison with a large range of alternatives (ideally, but a crazy ambition, every other possibility a person can conceive). As the series progresses the conception of probability changes, from something like ideal frequency in the early stages to something like degree of confidence in the later stages. The last stage is also the most demanding, in that it takes the most into account and requires an organized way of doing this.

The series has dangers. One danger is circularity. A false hypothesis can be confirmed because of its agreement with strongly held background theory. Another danger is bloat. Suppose that e would be explained both by a rich and complicated assumption H_1 and by a simpler more economical part of H_1, H_2. $P(e|H_1)$ will be the same as $P(e|H_2)$, and although $P(H_1)$ will not be greater than $P(H_2)$ they will not be very different because they are both novel and, let us say, there are no theoretical reasons to be suspicious of the added content in H_1 besides its irrelevance to the evidence. (In the terminology of chapter 1, H_2 will come closer to tracking the evidence than H_1 does.) As a result $P(H_1|e)$

69 Hawthorne (2005).

will be at most slightly less than $P(H_2|e)$.[70] The evidence does not discriminate them although it is the more economical theory that does the explanatory work. It is as if the mythical apple falling on Newton's head confirmed general relativity.

It might seem that comparative likelihood has the same problem, since the likelihood of the evidence given both theories will be the same. But there is a difference. The powerful H_1 and the economical H_2 are unlikely to be chosen as alternatives to one another, and the choice of alternatives is not itself a matter of comparative evidential support. The fact that it is not shows that the objective severity of a test, the topic of the next section, depends on many features of the context in which a test occurs on a particular occasion. This is an important difference between evaluating tests probabilistically and evaluating them objectively/modally.

Moreover if S is the additional content of H_1 over H_2 — H_1 is equivalent to H_2 & S — then the likelihood of E on H_1 is the same as its likelihood on H_1 & ~S, so that there is an alternative to H_1 that is not excluded, diminishing the support for H_1. Issues about bloat and circularity recur in the final chapter of this book.

There are many pre-scientific anticipations of taking prior probabilities into account. For just one example when evaluating the trustworthiness of two people who have both given rough predictions of an event that you have now observed, you consider both how well their predictions fit the event and how trustworthy you took them to be before this experience. It can take a lot of uncomfortable evidence to overcome the greater trust you have in some people than in others. But note that you would have taken less account

70 van Fraassen (1983).

of your earlier opinions if instead of saying "what do I think now?" you had framed your thinking as a test "let’s see who is more believable on this point". This is the greater weight that prior probabilities have at later stages of the sequence. There are also perceptual cases, as when you tell the optometrist "Looks like an A" although it seems to you most like the Cyrillic Я, which you do not expect to find on the eye chart. It is interesting that both examples, and others making the same point, concern the evaluation of sources of information. Their relevance, then, is more to the solidity or robustness of evidence than to the support that an item of that kind gives to a hypothesis.

severe tests

A hypothesis can pass a test, perhaps in contrast to an alternative, that is intuitively very easy to pass. It might be a significance test with a very permissive (high) confidence limit, or at the other extreme it might be a test considering prior probabilities of a hypothesis with very high probability in contrast to one with a very low probability So it would be attractive to take the severity of a test into account when considering how much support it gives to a hypothesis. But what is severity?

The sharpest and most influential contemporary discussions of severity are Deborah Mayo's. In a series of contributions she has honed and defended the idea that a statistical test should be hard to pass if the result is to count as evidence for a hypothesis. She argues that this is a unifying theme running through healthy applications of a number of otherwise different statistical methods and distinguishing them from

perverse applications. In fact, the subtitle of her latest book (Mayo 2018) is "how to get beyond the statistics wars". In that book she states a weaker and a stronger version of her motivating idea. On the weaker version severity is a necessary requirement for evidence

> One does not have evidence for a claim if nothing has been done to rule out ways the claim may be false.[71]

(I take the "nothing" here as rhetorical. We can understand the point to be that to the extent that little has been done that might lead to the rejection of the claim, evidence for it has not been produced.)

On the stronger version severity is also a sufficient condition:

> If [a claim] C passes a test that was highly capable of finding flaws or discrepancies from C, and yet none or few are found, then the passing result ... is evidence for C. [72]

Both requirements are very plausible and appealing. My concern is with the interpretation of "has been done" and “may be” in the weaker version and "capable" in the stronger version. The issue is counterfactual versus probabilistic formulations. Elsewhere, Mayo gives a rough formulation as “before regarding a passing result as genuine evidence for the correctness of a given claim or hypothesis H, it does not suffice to merely survive a test; such survival must be something that is very d!fIicult to achieve if in fact H deviates from what is truly the case”.[73] And she frequently gives counterfactual considerations to motivate her use of the statistical criteria. But the real work is done in terms of probabilities. Though she does not say this explicitly the strategy seems to be to understand possible situations in terms of the data produced in them and hence what the

71 Mayo (2018) page 5.
72 Mayo (2018) page 14.
73 Mayo (2005) page 97.

verdict a test would have had in that situation by plugging in that hypothetical data into the calculations that were actually used[74]. But the basic machinery is statistical, so the emphasis is on how often the same formal method would go right or wrong in other uses, rather than how easily it could have gone right or wrong in a particular situation[75].

There is a lot here for me to agree with, with the emphasis on subjecting hypotheses to tests which reveal restrictive conditions under which they would give wrong verdicts. I have some doubts about whether Mayo's apparatus does deliver what it is supposed to. My doubts centre on four topics.

(A) *the source of the probabilities.* This is the biggest worry. Suppose that they are wrong, and do not correspond to the frequencies that would show up in the long run and with many cases? (And which would then be evident in retrospect.) The most likely cause of this would be a misleading statistical model, based on a misapprehension about the type of process generating the data. Suppose for example that the experimenter takes the data to be governed by a binomial distribution when in fact the topic concerns a hypergeometric distribution, or even one that is a combination of the two, because replacement of individuals after trials is completely or somewhat restricted. But the experimenter has no way of knowing that this is the case, and similar systems follow the Normal pattern. This would lead to incorrect analyses of the data.

To deal with this worry while still keeping the actual force of evidence separate from the reactions of experimenters, one might impose conditions of predictive accuracy on the

74 This approach to counterfactuals has a lot in common with that of Judea Pearl. See Pearl (2000), chapter 7.

75 This is rarely made explicit in her writings. The nearest I know is excursion 6 of Mayo (2018). A telling footnote is number 2 on page 429, distancing herself from the standard semantics for counterfactuals.

probability assignments used in analysing an experiment. Instead of tackling the issue this way, in the next section I shall describe criteria that apply to the whole structure of a test.

(B) *Non-procedural features of the employment of a test.* Suppose that the very fact that the test was used is related to its truth? Suppose for example that a very tolerant type of test is habitually used by a very scrupulous and painstaking tribe of scientists, who have strong inhibitions against formulating and testing anything suggestive of loose conjecture. (They may have compensating methodological skills, or simply share a sharp nose for truth.[76]) When they apply their tolerant tests those that pass will usually be true. If a hypothesis is false then it is unlikely to be tested and pass. But this is not because of the virtues of the test but that of the scientists. (This point may be reminiscent of the issues about causal versus evidential decision theory. Also of the distinction between epistemic virtues and epistemic methods. It is related to the tangled question of causal inference, the topic of the next chapter.)

(C) *error/ignorance* We aim at a balance between error avoidance and ignorance avoidance, or the closely related distinction between type I and type II errors, because we have other aims for our inquiries besides accumulating knowledge for its own sake. But standard experimental method puts ideas through the gauntlet from preliminary evaluation to final exceptions with both types of consideration in mind. So the tests we are assessing for falsity-prevention are also the ones we rely on to give information we

76 A saying attributed to Quine runs "the universe is not the university", suggesting that the divisions between the different kinds of inquiry are more sociological than objective. No doubt this is frequently correct, but the possibility remains that in some cases an inexplicit tradition determining which kinds of conjecture are worth pursuing, based on an inductive awareness of what has succeeded and what has led nowhere, is accessible within a subject area and transmitted by training in it.

need in practice. And even if bare knowledge is the aim we may want results that will guide us in devising experiments that will get leverage on further questions. So one test might be less severe than another but more desirable. This would only happen with fairly severe tests and suggests that they might be *too* severe for other purposes. For this reason one might prefer to keep criteria of severity and criteria leading to theory acceptance separate. And one might be attracted to strategies like the metamorphosis one described earlier in this chapter, where different degrees of severity are suitable at different stages.

(D) *features of the experimental situation.* This is the smallest worry. Suppose that someone applies a very weak test to data that is in intuitively strong conformity with a hypothesis. (It might concern the tendencies of a random system such as a coin toss or the role of a tie or a roulette wheel, and they may be willing to endorse a conclusion that two thirds of the cases conformed to although on this occasion 99 out of 100 do.) Then the test result can provide only weak support for the hypothesis on a Mayo-like criterion although the data itself is strongly supportive.

The opposite is also possible: data of middling strength and an inappropriately severe test. Then a hypothesis could be rejected although there is at least enough evidence to move it through one of the early stages of evaluation.

These cases are neighbours of cases where a foible or irrationality of an experimenter leads them to accept or reject a hypothesis on the basis of an inadequate test. Then it

may be that a different experimenter would have got to a different conclusion on the basis of the same data.

Several themes run through these worries: the often non-ideal nature or situation of the experimenter, the possibility that an experiment may be badly constructed or interpreted, and the possibility of unknown errors lying behind the data collection. These are all "philosophical" in that they invite us to revise the borders of our terminology and to think hard about what intellectual device we use for what purpose.

comparative causal severity

The severity of tests is a matter of degree. Some are more severe than others, and this is especially important if we think that tuning them to stages towards acceptance is important. And in the present context it will be some sort of causal concept, relating to the conditions which would have led to an opposite result. But causation is not a directly quantitative notion.[77] This suggests building in gradations by defining a comparative concept where one test *as used on a particular occasion* is more stringent than another.

Analogues of Mayo's definition would feature relations between pairs of tests pairs of hypotheses, and evidence, where each test would choose one of the hypotheses on the basis of the evidence (alone), produced in a particular experiment and analysed by a particular person on a particular occasion using the probabilities that the person actually

77 Indirect quantitative aspects of causation are size of effect and probability of effect.

did use. The tests and hypotheses, however, are to be thought of as abstract objects, the ones that were actually used, and unchanged in other possible situations.

It is not hard to specify a relation of this kind that makes a stark contrast with a probabilistic definition and gives the intended results in many cases. Define test T_1 is *actually sharper* than test T_2 when one of them, T_i, was used and chose one of H_1 and H_2 over the other, using the probabilities that the person actually used and while T_1 put or would have put the two hypotheses in different pass/fail categories the other test either put them in the same category or reversed T_1's categorization, and the hypothesis that T_1 chose is nearer to actuality than the hypothesis that the other test chose or would have chosen (if used by that person with their probabilities on that occasion).

This is just one of a family of relations that can be defined starting with open sentences of the form "Ti puts H_1 and H_2 in different categories on the basis of e" and then combining them with modal and Boolean connectives and quantifiers while specifying which aspects are to be held constant over one possible situation to another and which to vary in accordance with the choices the person would have made. Actual sharpness, though, makes a very stark difference with probability-based criteria of stringency (although evaluating tests which use probabilities) and meets the worries that I expressed about Mayo's formulation. In particular it downgrades tests which would give wrong results — choosing hypotheses that are further from actuality than their alternatives — when used with the probabilities that the person is inclined to use.

The modal version also has the virtue of advertising the difficulty of knowing when it applies. It does not make the false presupposition that the probabilities one has used (and other crucial aspects of the test) are as they should be. On the other hand the modal version supports no such illusion. It is clear from its formulation that accurate probabilities used in a context where they will pick out the possibilities whose truth needs the least variation on the way things actually are.

The modal version also has transparent disadvantages. It is hard, sometimes impossible, to know which test has most actual discrimination. I do not think there is a definite concept of the stringency (severity, difficulty) of a test for a definition to capture. It depends on what you are going to use it for. Three important considerations about experimentation influence one another: balancing between error and ignorance, severity of test, and the placement of tests in sequences. How one wants to understand severity will depend in part on the balance between error and ignorance (type I/type II) one is aiming at and the way that one places particular trials in a sequence. All three considerations will be addressed again in the next two chapters.

chapter 5: cause with and without the help of experiment

This chapter is longer than others. That should not be surprising, given that I am presenting a causal account of evidence while discussions both of saying what causation is and of how we can support hypotheses about it are tangled and controversial. My discussion, moreover, covers a topic that is not much discussed, perhaps because it is though to be unproblematic: the connection between experiments as causal processes and inferences to causal conclusions. I shall argue that randomization plays a smaller role than is usually supposed in explaining why experiment is especially important in causal inference. The chapter divides into three parts: general considerations, accounts of causal inference, and ways of bypassing experiment.

Part I, some general considerations

It is hard to establish causation, and a controlled experiment is usually the most effective means. Consensus ends at that point. And rival accounts of the evidence for causation are rarely explicit about what their target is. The reason is that defining causation in a non-circular way is extremely, notoriously, difficult. Philosophical attempts to define causation usually reveal a range of causal relations of varying centrality. (And there are just a few claims that there is no such relation, or that it is scientifically or practically irrelevant.) I am not going to give anything like a survey of standard positions on confirming causal hypotheses. I am in no position to do that. Nor am I going to add to the definitions of causation.[78] I will however try to do something a little bit radical. The project of this chapter is to make a case for the suggestion that the variety of ways of extracting cause from data correspond to different aspects of causation itself, different

78 Allari and Russo (2014) is a wide-ranging and accessible treatment of accounts of causation.

ways of carving out coherent chunks from the large amorphous mass that can count as generally causal. These chunks have loose correlations with different uses of causal hypotheses. There is no single target for causal inference and no single way of performing it. One instance of this, what I call the depth of causation, will recur throughout the chapter. But I strongly suspect that there are other dimensions along which similar points could be made.[79] The picture can be put together in terms of the development of causal hypotheses from claims about the direction and overall structure of the causal links between several observable quantities to claims about the detailed processes leading from one quantity to another, or to claims about the balance between influences in different contexts.

Begin, for once, with pre-scientific belief. We have many opinions about what causes what in everyday life, and a few of them are true. (An irony: causation has all the marks of a common sense concept, needing refinement and likely division into more real components before it is scientifically useful. But common sense is pretty inaccurate in attributing it.) We arrive at many of them by the following method. We begin with broad classes of pairs of events where producing one results in the other. If we want to get the second then doing the first is often effective.[80] (There are causal relations between individual events — token causation, sometimes also called event causation — and between kinds of events — type causation. The relation between the two is discussed at the very end of this chapter.)

79 Three related candidates, each of which comes in degrees, are the degree of causation – Braham and van Hees (2009) - the probability of the effect, and the size of the effect.

80 Woodward (2016 [a]).

A warning is needed. Pairs of events that we see to be effects of a common cause, often with scientific hindsight, can guide effective ways of acting, in normal circumstances. You are a terrified passenger trying to control an airplane whose pilot has collapsed. You want to maintain your distance from the ground and visibility is not good so you rely on the altimeter. It works. You act as if maintaining the altimeter level causes the altitude to stay constant. But it doesn't *really*. If you reset the altimeter this would have no effect on your altitude. It's the other way around: constant altitude causes constant readings, via unchanging atmospheric pressure. The noteworthy thing is not that we can be confused about causation but that we can apply these backwards relationships successfully, even when we know they are inaccurate. And for what comes later this chapter it is important that understanding which event is a cause of which and which events have common causes matters most when the connection between them is variable. If you are flying on an irregular boundary between the desert and the ocean air pressure will change abruptly so that the policy of the terrified passenger may misfire.

Sometimes, too, we make something happen by controlling a class of events that is linked to what we want to achieve by the presence of a common cause. You want to make someone smile so you act in a way that makes them laugh. But neither the laughter nor the smile is the cause of the other. They are common effects of happiness. But you cannot know anything directly in terms of happiness so you guide your actions in terms of a consequence of it in order to effect a different consequence.

depth of causation, causes versus conditions

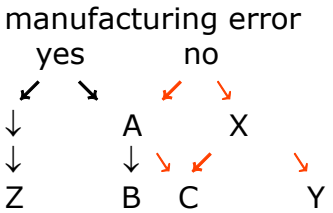
There is a basic reason why actions do not need to be based on means-to-ends causation. Causation comes in various depths. Given that altitude and air pressure are related as they are at the time, changing the altimeter reading will be accompanied by change of altitude. But in a situation in which this background is not constant the connection does not hold. Then factors that change or could change the correlation have to be taken into account. The link between the altimeter reading and air pressure is *deeper* than that between altitude and the altimeter reading. It produces a situation where the correlation can seem like one way causation, and for some purposes can be successfully taken as causation. There are many examples, often not as transparent as this one.

Depth of causation will play an important role in this chapter. The background is the distinction between more and less fundamental laws (principles, regularities) of nature. Unsupported objects on earth as it is accelerate downwards at 9.8 m/sec^2, but this is a consequence of fundamental facts about gravitation, which would be so even if the mass of the earth were different. (We expect that these gravitational facts are themselves consequences of yet more fundamental matters, though we do not understand what they are.) The deeper principles underlying any scientific field that can be reduced to a more basic one will give further examples.

More and less fundamental natural processes generate possibilities that are more and less easily obtained, nearer and further from actuality. That ball free-falling from the tower of Pisa at 9.8m/sec^2 is a near possibility, even if it does not happen. (All it takes is for Galileo to change his mind.) On the other hand an object's falling at 15m/sec^2 is a

much more distant possibility, requiring a different history in the formation of the earth where it becomes more massive. Counterfactuals exploit the nearest possibilities where their antecedents hold, so although they are intelligible with remote antecedents, when the antecedent can hold in both nearer and more distant situations the counterfactual will opt for the nearer one. If there had been an object falling at that spot at that time it would have reached the ground in accordance with the 9.8m/sec^2 rule.

Causation can also exploit both nearer and more remote possibilities. The formation of the earth is a distant cause of the fall of a particular ball now. Things get complicated, though, when a cause can happen with different outcomes in both an easily realized and a more remote way. Consider an example. Due to a defect in its manufacture one leg supporting a pinball machine is slightly longer than one on the other side, leading the machine to tilt slightly to the right. A ball is shot into the machine and strikes bumper A. As a result it strikes bumper B. But if the defect had not occurred the ball would have struck A at a different angle and so would instead gone onto to strike C. So the manufacturing error is a distant cause of the ball's striking B. It causes the collision with A to cause the collision with B and if it had not occurred the A collision would not have caused the B collision. The moral is that going back a stage in the chain of causation can reorganize the causal connections from that point on.[81] (Diagram below, where paths exclusive to the second trajectory are in red.))



81 Someone might object that the situation cannot really occur. But pinballs *do* take different paths depending on where they come from before striking a bumper.

There are familiar cases of this phenomenon. This application of the antibiotic causes the recovery of health. But if the antibiotic had been applied too often it would not have been a cause of a recovery. The influence that changes existing causal patterns need not be a variation on an actual cause. The presence of the virus causes the fever, but if an inoculation had occurred it would not have. Intuitively deeper causes are less susceptible to this kind of reversal.

In this example there is an element of indeterminism (which can be due simply to the granularity of the description) which leads to another feature. Following the paths further we see that A is a (potential) cause of D and the Lewisian not-not characteristic applies — if not A then not D — on the nearer paths. But this is not so on the less near ones (in red). This additional sign that including deeper causes can scramble the arrangement of connections may also ground a suspicion that there are several connections between events going under the loose label of "causation".[82]

The relation between X, Y, and C illustrates an important point. When additional causes are introduced quantities that would have had an unmotivated correlation (C and Y) before the expanded picture are now seen to be effects of a common cause (X).

Statistics collected under non-experimental conditions will nearly always reflect surface causes of the phenomena. What else are they likely to, given that these are the normal

82 It would be remarkable if causation remained a univocal matter when understood scientifically. Very few commonsense concepts do.

influences? (You do not compile data on downward terrestrial acceleration by imposing a variety of masses on the earth.) This creates differences in both aim and data between purely observational and experimental research.

The events that I am classifying as deep causes are closely related to the conditions under which a surface cause can have its normal effects. Releasing an object near the surface of the earth causes it to accelerate downwards at the standard rate, *given that* the earth has the mass that it has. But also striking a match causes it to light given that there is oxygen and not too much humidity around. Such conditions are nearly always necessary but not sufficient and nearly always have formed before the time that the cause operates. Sometimes two conditions serve to maintain one another. Then there is an equilibrium between them.[83] Examples are the quantities related by the gas laws, by Hooke's law connecting the extension and force of a spring, and those which are stable in physiological homeostasis. Or for that matter the affection between two people. Neither is a cause of the other in any sense that is exclusively one way.[84] These will return later in this chapter.

Issues about depth of causation and issues about evidence come together when a correlation that is not directly causal gives support to a hypothesis. An example is when someone's having a high temperature is evidence that they will show the other symptoms of the flu. But neither temperature nor flu symptoms cause the other. Instead they share a common cause, exposure to the influenza virus. This is so with the relatively deep cause of viral exposure. If instead we think in terms of the shallower cause of

83 Wilson (2017) suggests that the contrast between causal and equilibrium quantities corresponds to that between elliptic and hyperbolic differential equations. The connection with succession versus simultaneity is clear; that between surface and substratum is not.

84 In this connection see Andersen (2013).

having the flu, as we might when considering a single patient who we know has this disease, then temperature is evidence of the disease and is caused by it. The general phenomenon is that we take facts that are causes or effects of a type of event as evidence for it, but since such direct causation often disappears when deeper causes are considered the evidential connection then turns into a reliable correlation under current conditions. It is a weaker connection, though, since having the flu accompanies a fever with more exceptions than exposure to a flu virus does. This makes sense on an objective understanding of evidence, since in terms of having a disease there is a more fragile basis for the conditional-like relations that constitute knowledge and weaker analogues of it than in terms of exposure to the infectious agent. I take this as support for the objective account.

Mill and simple control

John Stuart Mill's application of his methods of similarity and difference to determining cause, in *A System of Logic,* is a precursor to contemporary sophisticated accounts of causal inference. Mill is trying to extend his account of inductive reasoning to an explanation of the power of experiment over cause in particular. In the crucial passage he writes.

> When we can produce a phenomenon artificially, we can take it, as it were, home with us, and observe it in the midst of circumstances with which in all other respects we are accurately acquainted. If we desire to know what are the effects of the cause A, and are able to produce A by means at our disposal, we can generally determine at our own discretion, so far as is compatible with the nature of the phenomenon A, the whole of the circumstances which shall be present along with it: and thus,

> knowing exactly the simultaneous state of everything else which is within the reach of A's influence, we have only to observe what alteration is made in that state by the presence of A.[85]

Mill is clearly right about one thing. When we can produce an event we can often control when and how we do so, so we are in a position to produce observations about what else is present with its effects, and to that extent we have a combination of factors that will normally produce the effects. But behind this truth there are a number of important further points.

The simplest problem is that it ignores the inevitable statistical dimension. One is supposed to create the cause a few times under varied or rarefied conditions and see what happens, noticing what is always or never there. But it is never really this neat. There will be exceptions and intruders even when the true causal pair is tried. So one needs to consider averages, curve fitting, and the like, and the all-important choice of a sample. Then non-obvious ways of extracting a meaning enter. The meaning is rarely conclusive or unambiguous.

One reason for this is that there are normally factors that one cannot control, has not anticipated, or simply does not know about. What I referred to as insulation in chapter 2 gives a measure, but only that, of protection against these. (And one can only insulate against what can be insulated against.) These confounding factors will have a greater presence on some occasions of an experiment than others.

85 Mill (1843) chapter VII, section 3.

A related problem is the lack of a clear and definite verdict on the direction of causation. Suppose that the experiment produces A (turning on light A with switch S, say) to see if it brings about B (light B illuminating) and B is then observed. A may cause B (the circuit goes from S to A, and another switch sensitive to the light from A, turns on B). But it may also be that the experiment has produced A only by producing B. (The circuit goes to B, sensitivity to whose light turns on A.) Or it may have produced some common cause C which has led to both A and B (S is linked to a hidden light C, to which both A and B respond). The solution is to record the incidence of B without A. If such cases are found, in sufficient numbers, then it is A that causes B rather than the other way round. This would rule out a common cause also. (The sufficient numbers proviso is necessary, because the causal connections could have exceptions or be easily blocked by intervening extraneous factors.) In the ideal case one could infallibly produce A which would frequently lead to B, but B was either impossible to produce except by producing A, or there was an infallible way of producing B and it fairly often was not accompanied by A. (Strangely, this is most effective when there is a little noise in the system and the causality occasionally fails.) In the absence of such data the direction of causation has to be decided on the basis of temporal order, which does not always apply, or background theory making processes leading in one direction more likely than in another, which may not exist.

Call this the "make it yourself" strategy. It fits necessary parts of sufficient conditions in familiar situations — properties of a system which when removed make consequences no longer apply as long as you do not change things very much. These are captured by Mackie's classic INUS formulation. In Mackie's somewhat enigmatic formulation a cause

is an insufficient but necessary part of an unnecessary but sufficient condition for the effect.[86]

The make it yourself strategy defines a relation between classes of things or events. Sometimes this is all the causal depth that we need. You are treating an infection with an over-the-counter medication. You know that the infection is microbial in origin, and that it is the reason that you have a high-temperature. So you take your temperature and swallow the medication and a couple of hours later take your temperature again. Suppose it has declined. You know that degree of infection and increase in temperature are correlated, so that the reduction in temperature is a strong indicator that the medication is having an effect. Next time you have an infection like this you will take it more readily. You do not need to know what the detailed connection between the medication and the infection is. It is enough that there is a connection, and that using it to choose a treatment is effective.

There are many cases where this is not enough, cases where the direction of causation matters and time or more background is not going to determine it. You notice that when you have a sore throat have a headache; should you take a throat lozenge or a Tylenol? It might be that either causes the other or that they are both effects of a virus. The centrepiece of experiment-based ways of tackling such unguided causal issues is randomization.

86 Mackie (1965). Mackie refers to Mill several times, obviously having learned what is helpful and what is fragile there.

randomized application

It will not be news to many that the standard scientific technique is now randomization, with the hope of evading these problems. Nearer to being newsworthy is that randomization has ambiguities and limits, that there is some dispute about why it is effective, that there are precedents for it in pre-scientific life, and that there are aspects of causation that it does not handle well. These illuminate one another.

Randomized application of a candidate cause (distinct from randomized selection of a treatment group, discussed in chapter 3) begins with applying the candidate to a randomly produced selection from the population of interest. The statistics of the result are compared to what happens in a control group which has not been exposed to it.[87] (In a variation, two possible causes are used and compared.) The idea is that, at least in an unattainable idealized case, the application of the possible cause, or the determination of which potential cause is applied, is correlated with nothing else, so that any correlations with the possible effect can have no other basis.

This requires the experimenter to be completely in control of whether the treatment has been applied. (Or at any rate to know with certainty when an attempt at applying it has succeeded or failed.) It requires also that the experimenter know that nothing other than experimentally applying it will result in its presence, in the context of the experiment. I have the impression that this second requirement is not often explicit. It is needed because otherwise there could be other causal influences, particularly in the opposite

87 Fisher (1935), chs 9, 10, Lehmann and Romano (2005) ch 5, Wasserman (2004), Bennett (2013), Ruxton and Colegrave (2003 chapter 3).

direction, from the effect to the application, which would give the illusion that the application had caused the effect in this case.

Situations where the application of the cause is subject *only* to the random process are unattainable and idealized because there are (nearly) always other determinants of the application. We inject a medication but its absorption depends on features of the patient's metabolism and the technique of injection. We try to keep these as uniform between the two groups as possible, but we can never completely succeed and we can never be sure how near to complete success we have got. As a result the causal aspect of the conclusion is blind to many things. This is a greater danger in the face of additional factors that we do not know about, because we cannot compensate for anticipated limitations of randomization with respect to the hypotheses and subject matter concerned by designing the experiment to obstruct or control them. Or, to put it more carefully, any such compensation will rely very heavily on background assumptions about the kinds of courses likely to be operative, and the more interesting or original the hypotheses the greater the risk that these will introduce errors. So the comparison is always implicitly between treatments under vaguely specified conditions, including the "treatment" where one does nothing artificial.

There are commonsense analogues to randomization. One can apply the conjectured cause in an irregular way, as unlike a law-governed order as one can make it, and see whether the conjectured effect follows in the same pattern. So for example one could see whether a light is controlled by a switch by flipping the switch at intervals spaced according to the digits of π, or the like. Then one would check whether the light

illuminated in the same way, perhaps according to some simple transformation of the input pattern. If the light rarely goes on except when the switch is on, and one knows that the switch comes on only by one's deliberately activating it, then the conclusion that the switch controls the light is strongly suggested.

For very strict Bayesians, for whom probabilities are simply degrees of belief and thus are not coupled to causes and effects in their objects, there is less of a direct reason why randomized application should work. The assignment of individuals to the two classes can still be unpredictable and not correlated with various extraneous factors, by manufacturing a situation in which an unpredictable device such as a table of random numbers is the sole factor in assigning individuals to the two groups. Then subjective probabilities will carry no information about the assignments. In effect, this is the approach of "objective" Bayesians, whose main concern is to legitimize the use of prior probabilities for hypotheses and often data. The question remains why hard core Bayesians should have anything to do with causation in the first place, believing as they do that everything relevant must be expressed in terms of degrees of belief.[88]

It is important to note that the treatments or other suspected causes still have to be applied physically. The application still has to be causal. Pills have to be presented and swallowed, injections injected, chemicals mixed, heat applied. Once an individual is given its random assignment that assignment still has to be carried out. If this were not so something independent of the assignment would be the physical cause of the treatment and this would not be excluded as a possible cause of the individuals exhibiting one effect

88 Sophisticated versions of the idea that causation is increase in probability are not decisively refuted. But the fit is not at all plausible when the probability in question is an attribute of a person.

rather than the other. Suppose that we assigned individuals to one or the other group in accordance with their colour, and then noticed that the pattern of assignments happened to coincide with a table of random numbers. That would not count as random assignment in the relevant sense. The randomness has to be part of a process that physically makes one treatment or the other occur. Or suppose that we assigned objects to groups in some arbitrary way, and simultaneously we activated a mechanism which searched through a table of random members until it found one that corresponded to the assignment, then presented this to the experimenters as the random assignment of the experiment. We could arrange this so that the probability of any set of individuals was proportional to its size, just like a random selection. But this too would not serve the purposes of real active randomization either. Purely statistical considerations will not do it: there is no way of magicking a causal fact out of facts about proportions and limits.[89]

A consequence is that a randomizing act or process must complete and mingle with other causal influences on the observed phenomena. So its influence-limiting power is not absolute. To stick with the standard example, if a drug is given by injection to a random selection of individuals the influence-monopolizing power of the selection is still constrained by the fact that it combines and presupposes other factors. Unexpected influences may sneak through in terms of their connections with the performance and perception of the injection. Suppose that the experiment is supposed to show that the active ingredient of the drug is a unidirectional cause of some aspect of the health of the individuals. However there still can be common causes of that aspect and absorption or metabolization of the ingredient. This could be clinically significant. Randomization, then,

89 Contrast this with the abstract version of Fisher's conception in Greenland (2011). The overall topic is causal inference but the necessary conditions for instantiating randomization to get real causal conclusions are left implicit.

is not a magic way of discriminating cause from correlation but a source of fallible evidence about which is present, to be weighed against evidence from other sources.

part II: causal inference

I will compare four ways of getting conclusions about causation from data. Three of the four are are meant to be used in conjunction with some input from experimental data.

from individuals to classes and partially back, potential outcomes

The potential outcomes approach has been and possibly still is the dominant approach to causal inference in applied statistics.[90] As with all standard techniques from statistics for getting to causal conclusions, the target is a relation between types of factors or attributes along the lines of "A is one of the causal reasons why B occurs". But it does this on a detour through causal connections between particular events. There are two core ideas. The first is is that if we have a sample from a population and if we knew of each individual in it both whether if it were C it would be E (which is immediate, just part of the data, if in fact it is C) and whether if it were C it would be not-E (which is immediate if in fact it is not-E), then we would have evidence whether or not being C is a cause of being E. It would be real evidence to the extent that the sample is typical, and chosen in a way that does is independent of the possible causes to be investigated, both of which are encouraged by making the selection random.[91] The second core idea is that

90 The basic suggestions come from Neyman (1924/1990) and Rubin (1974). A compendious exposition is Imbens and Rubin (2015). A simple though limited exposition is chapter 16 of Wasserman (2004).

91 When $\rightarrow$ is the ordinary English counterfactual as formalized in the dominant Lewis-Stalnaker way, and $\supset$ is the material conditional this is equivalent to $((c \supset (\sim c \rightarrow \sim e)) \ \& \ (\sim c \supset (c \rightarrow e))$, where c is the conditions under which the sample is taken. I am sure that taking them this way and confronting the formula with intuitions about examples it would emerge as both too weak as too strong.

we can get a good grasp of whether an individual would have had one attribute were another attribute to apply to it by comparing it to others that are similar to it in other respects and have been exposed to the possible cause. Similar things will behave similarly. If other people of the same gender born in the same year and with the same medical and social histories — or different stages of the same person — were E when they were C, then it is a good bet that a person who never happened to be C would have been E if they had been in that position. This section describes just enough about the technique to fit it into the themes of this chapter. But it has been developed into something much more elaborate and comprehensive.[92]

The aim is to extract a verdict about causation suggested by these two ideas. The criterion is directed at situations where there is a treatment and a desired effect, and one standard formulation compares averages of individuals in the sample with and without the effect following the treatment. (Others are more explicitly probabilistic or work in terms of ratios instead, but the effect is the same.) Consider the table below (which because of the argument which follows has more categories than is usual).

	T observed	E observed	potential T	potential E
A	yes	yes		
B	yes	no		
C	no	no		
D			yes	yes
E			yes	no
F	no	yes		
G			no	yes

92 Dawid et. al. (2019) associates this idea with the difference between investigating whether one factor is a cause of another and investigating whether a factor causes a known effects backspace. While that distinction not my immediate object it coheres with a theme of this chapter. This too was anticipated by Mill. Mill (1843) book III, ch.X sec. 7.

Potential **T**reatment and **E**ffect here mean that the individual would have presented the corresponding observation if exposed to the relevant condition, in accordance with the second idea. Then the sample gives evidence that the treatment is a cause of the effect when there are more in category A and category D in the sample than there are in B and D. The sample suggests that the treatment is not a cause of of the effect if these sizes are reversed. On the other hand it suggests that the effect is a cause of the treatment when there are more in A and B than there are in F and G. And the suggestion is that the effect is not a cause of the treatment when this ordering is reversed.[93] So there are resources here for distinguishing between directions of causation. And we can take as evidence that there is a common cause at work when either both of the negative possibilities are false or all four ordering-suggestions fail.[94] (The latter disjunct would occur very rarely, since even when there is a common cause noise in the sample is likely to prevent strict equality.)

The distinctive feature of this approach is also its weakness. The vulnerability lies in extrapolating from the data in an actual sample to data that would have been presented if all the sampled objects had been exposed to the treatment and then observed. The most worrying element of this is the use of some actual objects as surrogates for others.

Which objects are good comparators for which others, in terms of how they would react to a treatment, is surely a contextual matter. A comparison between rats and people can be causally enlightening when the topic is the effect of drugs but not when it is reactions

93 These are usually expressed in terms of averages, but since all the samples here are the same size sums are simpler.

94 When both positive ordering suggestions are true one possible interpretation is that T and E are in equilibrium, as mentioned earlier in this chapter. States in equilibrium are usually both effects of a common cause and causes of one another. Examples where both are true but the physical situation seems not one of equilibrium would be bad news for the account.

to advertising. So the causal attributions that result from potential outcomes would best be labelled as causation relevant to a purpose. Such a labelling would have definite uses, but it suggests that the most suitable place for results we get by this method is at some particular point of theory-development and application.

One special case is particularly enlightening. Experiments are usually conducted under artificial and often unnatural conditions. That is one reason why they can be revealing about the deeper causes of events. But the selection of subjects for an experiment is usually done outside the conditions of that experiment. Moreover good experimental method takes account of unknown influencing or interfering factors as well as known ones, and these are not available when paring objects up as representatives of each other's potentialities. That makes a mismatch. The technique was intended to be used when interventional experiment is not possible, and we now see reasons why its results may not be reliable within an experiment. The moral is that it is best confined to non-experimental research. But there are limitations to the parts of nature that this can reach.

There is a definite resemblance here to theories of causation in terms of counterfactual conditionals. The statistical use of material conditionals — "most" — is analogous to the "variably strict" aspect of subjunctives — their dependence on what happens in a nearest possible world. And the use of pairs of objects, only one of which has both been exposed to the treatment and the presence or absence of the effect observed, is analogous to once-lively discussions of counterparts in other possible worlds of actual individuals. On the one hand the method is based on a way of simulating what would have happened in

a particular case. But on the other hand its results do not entail conclusions about particular cases. The most we could get would be probabilities. But there would be counter-intuitive results. The probabilities would be for claims saying "if e had occurred then f would have" where e and f are particular events. These would be a bit unorthodox when e and f did actually occur. And they would run afoul of the large body of counterexamples that have accumulated for simple forms of Lewis's not-not analysis, concerning preemption, over-determination, and the like.[95] I doubt that practitioners of the method would be very bothered by these. The focus is on causal relations between types of event, and these are largely unaffected by any peculiar consequences for token events. On the other hand too many such consequences would suggest that the types are not causally fundamental, not the groupings that would be found in the laws of nature, or alternatively that the language used for picking out individuals that can model one another is inappropriate. In either case would give a warning sign that the method had been applied in a potentially misleading way.

To sum this up, the factors that are both strengths and weaknesses of potential outcomes are its inflexibility in the face of conditions of varying severity and its implicit use of purpose-relative causal substitutions between different objects.

from cause plus correlation to more cause, causal modelling

Another technique for squeezing causal conclusions out of data is based on graphical causal models. Causal modelling does not extract causal conclusions from purely statistical data, except in some special cases, and in fact it is often accompanied by an

95 Straightforwardly in Lewis (1986), and with complications in Lewis (1996).

insistence on the irreplaceability of causal thinking. Instead, the aim is to extend some causal knowledge further, combining assumptions about causation in some domain with data from the domain to give further causal conclusions.

Graphical causal models are familiar to most philosophers interested in causation and many statisticians, especially those who appreciate the difficulty of getting causal conclusions from purely statistical (proportional) evidence. I shall not give a detailed exposition as some features need a lot of care and some formalism to state accurately, though they will be familiar to some readers and frustrate others. However excellent expositions are available.[96] They give a way of representing a set of variables representing objective quantities that can take a fixed range of values, causal relations between these variables, represented graphically by arrows, and the degrees to which variables affect one another, represented by functions from the range of one variable into that of another.[97] These come with a probability distribution that determines the probability of variables' taking particular values in terms of the values that other variables upstream along the arrows can take, and ultimately in terms of "exogenous" inputs from processes that are not part of the model. There are a number of conditions that the arrows, the functions, and the probability distribution must satisfy. The two simplest are that there must be no closed cycles in the pattern and that the probability of any variable taking a value is a function of the variables immediately upstream from it. The technique comes with the assumption that causal relations between quantities can be represented by models of this kind.

96 Chapters 1 and 2 of Pearl (2000), Pearl, Glymour and Jewell (2016).

97 Of course the intellectual content of the technique could be separated from the diagrams. But in practice this is never done

For present purposes, the important point is that if we are given a partial specification of a causal model — some and sometimes none of the arrows and some or usually all of the probabilities — then there will be a limited range of ways the model can be completed to make a model satisfying the conditions.[98] In particular some ways of filling in the causal connections — putting heads on the arrows, turning them from bare correlations into directed causal relations — will be incompatible with the probabilities, given the structure that a causal model is taken to have. For example, if two variables A and B are both correlated probabilistically with a third variable C but not correlated with each other, then C cannot be a cause of A and B, since descendants of a variable (variables causally downstream from it) are correlated with it. Many of the restrictions on ways of filling in a partial specification are, like this, ways of determining that one variable is *not* a cause of another. They typically work by predicting correlations that would hold if one variable were a cause of another; then if the correlation is not found the causal relation does not hold. As a result they usually concern the opposite end of the graph from interventional considerations, downstream rather than upstream. (The statistics themselves have to be derived from samples in any of the usual ways. Notice the necessity in this of statistical models, which as I have argued involve implicit causal assumptions.) There are some surprising and subtle examples of the result. For example, the relation between smoking and tar accumulation can help confirm that smoking is a cause of cancer.[99] The more interrelated factors are in play, the greater the clarity about their causal connections.[100]

98 Glymour and Cooper (1999)

99 Pearl, Glymour, and Jewell 2016, chapter 3, Pearl 2000, epilogue.

100 The process can get intricate; there is a computer program: Spirtes, Glymour, and Scheines (2001). (An obstacle here is that the current model searching software will not handle a variety of distributions.)

This is a novel and interesting approach, breaking with traditions in both statistics and philosophy; a fruitful middle layer of analysis between two phases of experimental research, a nice empirical/conceptual/empirical sandwich.[101] I have sniping criticisms and a fundamental point.

Sniping: the graph-refining routines will not count any link as causal without a correlation of their values; more generally they rely fundamentally on which variables are and are not correlated. This will generate a tendency to ignore very slight causal connections, unless the samples are very large, because minor correlations will often not show up in them. But a correlation that is too small to emerge from the statistics as significant in a limited domain may be much larger in an expanded domain, as the pinball example shows.[102] And a correlation extracted by a suitably designed experiment may not be apparent from non-experimental observation. This may diminish the force of the technique as a substitute for experiment.

This sets up my greatest concern about graphical causal models. Some experiments reveal causal facts concerning deeper layers of causation than can be produced without characteristically experimental measures. These will be invisible using the non-experimental statistics that are essential to causal modelling. So some tension between the output of the enterprise and the results of experiment is inevitable. That means that the modelling technique will not be effective in basic physics or other sophisticated sciences. To my mind this seriously diminishes its attractiveness.

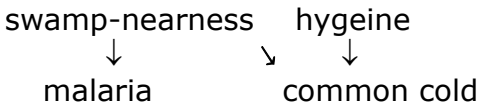
---

101 There are other ways besides the causal models approach of carrying out this general program. See Gopnik and others (2004). I shall not discuss them.

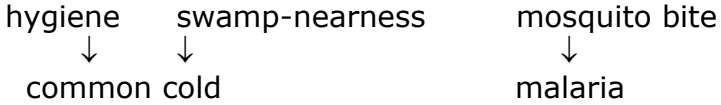
102 This is related to the freewheeling attack on the whole program in Freedman and Humphreys (1999).)

The problem also has an impact on characterizing causes in terms of intervention.[103] If the intervention does not involve insulation and the like then the causes it picks out will be shallow ones, useful for every day choice of actions and descriptions in everyday life. On the other hand if it does involve these things then it will not pick out the ordinary targets. Since the statistics needed for defining the correlations are likely to be observational, the causes will be shallow, and only hints of underlying processes. They are likely to reveal the causal patterns that appear when causes are interacting and swamping one another, rather than those that are brought out by insulation and randomization.

A two-stage example of this (although the pinball machine example earlier could illustrate the point): getting malaria is correlated with living near standing water. Refer to all standing water as swamp. If this is all we had to go by then it will be reasonable to count swamp-nearness as a cause of the disease. So we will suppose that the causes look like this:

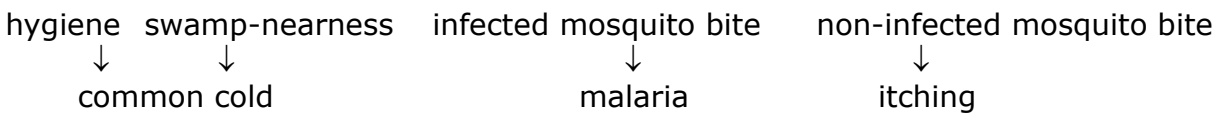


A very rudimentary experiment subdivides swamp-nearness into with-mosquito-bite and without-mosquito-bite. So we get the following, where swamp-nearness is not a cause of malaria:



103 As in Woodward (2003) (2006) (2013) (2016).

A more sophisticated experiment will show that only the bite of mosquitoes infected with the *Plasmodium* parasite is correlated with getting malaria. So we get the following, where getting bitten is not by itself a cause:



Adding antecedent causes rearranges the connections between factors. But now one particular way this can happen emerges. The acquisition of a new concept, which is often though certainly not exclusively the result of what experiment reveals, can bypass factors that were thought to be causes. I would not regard any account of the epistemology of causation as satisfactory unless it made space for this possibility.

After pointing out essentially this feature of their approach, in connection with a treatment of token causation ("actual causation" in their terminology), Pearl and Howarth write:

> ... the truth of every claim must be evaluated relative to a particular model of the world; that is, our definition allows us to claim only that C causes E in a (particular context in a) particular structural model. It is possible to construct two closely related structural models such that C causes E in one and C does not cause E in the other. Among other things, the modeler must decide which variables (events) to reason about and which to leave in the background. We view this as a feature of our model, not a bug. It moves the question of actual causality to the right arena—debating which of two (or more) models of the world is a better representation of those aspects of the world that one wishes to capture and reason about.[104]

104 Howarth and Pearl (2005)

This backs up my point. It points to the best application of the method, discussed below. All I would add to it is a) that embedding models in larger models tends to turn causal relations into effects of common causes, and b) that taking causal modelling as having an intrinsically provisional quality undercuts some claims that it will give mechanical ways of detecting causation in data.

Closely related to the worry about depth of causation is one about the homogeneity of insulation and other controls. Suppose we have constructed a causal model on the basis of statistics gathered by several distinct experiments, discerning the correlations of different variables that the model aims to find causal connections between. Suppose also that these experiments involve controlling for different possible irrelevant factors with different degrees of effectiveness. Then there will not be any uniform level of cause-induced-correlation to which the arrows in the model correspond. Such a model is not likely to give reliable results in practice.

stages towards causal conclusions

We begin by wondering whether two quantities are causally related, and if so how. So we begin by collecting statistics, and apply tests for whether any correlations they suggest are significant. Then really rudimentary criteria of causation such as temporal order, and well established textbook-style theory will reduce the causal possibilities. These can be further refined with more structural considerations, such as those provided by graphical causal models. This is likely to be most effective at a very early stage where very minimal controls encourage the various sources of the data to be causally homogeneous. With a refined rough draft given in this way we can proceed to compare pairs of

hypotheses, both with GCMs and with instrumental-variable-augmented potential outcomes. GCMs will have a particular role in filling in causal processes between the gaps of individual causal relata. They cannot be self-sufficient in this, though, because this task relies on a choice of appropriate concepts which usually have to come from fundamental theory.[105] And evidence about which causal connections will be most significant in which contexts, for example relevance an intended application or an intended medical context,[106] will usually need particular experimental controls — insulation, choice of trigger and sample, focus of randomization — suitable to those contexts. Often at the end we have a theory of processes and mechanisms from which we can get causal conclusions but which does not in itself employ the concept of causality. It may even have anti-causal implications.

At the different stages of this typical development of a causal hypothesis the relation between type events and token events as causal relata usually changes. Distinguish three stages of causal theorizing. The first begins with supposed connections between individual events and very small classes of events, trying to piece them together to get means to a constrained list of ends. So the focus is on two-way induction between types and tokens where types are small collections of tokens. The aim is to get accurate predictions of particular token cause/effect pairs. In a second stage we consider larger types of potential causes and effects, typically against a background of particular not too ambitious but roughly exception-free causal relations between small classes of particular event. Reconciled to some degree of exception at the token level between causally paired types. So a conclusion that one type causes another is not taken as entailing of any

105 Some processes need generalizations of the method that do not yet exist, though. For they concern continuous linkages between cause and effect.

106 Cartwright (2004), (2010)

particular events belonging to the one that it will belong to the other. The aim is to estimate the typical or average influence that one factor will have on others, as described in the quotation from Pearl above. Event types are then naturally understood as broad combinations of scientifically significant properties and physically concrete individuals.

At these first two stages the emphasis is primarily on *whether* the type-to-type causal relation holds between particular event types – although the interpretation of event types in terms of event tokens is different at each. At a third stage it shifts to *how* the causal connection is made. This largely consists in breaking the role causal connections down into more detailed causal chains — processes, mechanisms — which can be understood in terms of related physical theory. As suggested above causal language may drop out at this point. So too may reliance on any kind of event type, since many of the purposes of primitive causation can be served for the relevant kinds of event by appeal to chains are influence understood in the appropriate way. (For a trivial example the motion of one solid body upon impact from another can be predicted and explained in terms of the conservation of momentum, if necessary extended and combined to apply to complex bodies.)

The acyclic requirement, that there not be closed causal paths from an event eventually back to itself, means different things at these three stages. At the first stage it can be taken as a very simple constraints on patterns of causal relation between token events: they cannot go around in loops. This can not be the interpretation at the second stage, although it is best known here. For its main application is to rule out unwanted causal patterns between correlated types of event. And individual links of the resulting causal networks will very often relate types of event for which some pairs have a cause/effect direction opposite to that of the overall network. (Injection of an antibiotic generally causes reduction of an infection, but particular cases of fever reduction can be among the causes of injection, for example they increase confidence that the disease is properly dealt with in this way.) So the constraint must mean something like "adequate estimates of the average effects of causes, with the emphases and inclusions required for the purposes of the particular model being constructed, must be possible without including factors causally antecedent to a variable in the estimate of its effects". This will certainly not be satisfied by type events that are simple collections of all qualifying tokens, illustrating the point that the evidential assessment appropriate to a stage of causal theorizing shapes the implicit ontology of the resultant hypotheses.

Part III: doing without experiment

instrumental variables and natural experiments

Sometimes the data available to us has *some* of the power of experiment although an experimental intervention would be impossible, or unethical. I shall first discuss the use

of instrumental variables and the related topic of natural experiments, and then the rather different case of astronomy.

In the instrumental variables technique we investigate whether one quantity is a cause of another by finding a third quantity about which it is uncontroversial that it has a causal relation to one of the two, could cause the other only through the first, and is caused by neither.[107] We can compare lung cancer rates in locations where cigarette taxes are at different levels, on the assumption that taxation levels have no influence on lung cancer and that people smoke (somewhat) less if it is more expensive.[108] If we find less lung cancer where cigarette taxes are higher then that is evidence for a causal connection. In similar ways draft lottery numbers for US conscription, postal (zip) codes, and month of birth have been used as instrumental variables in suitable contexts The causal relation is often incomplete and partial, so that quite sophisticated data analysis is needed.[109]

The use of month of birth and postal codes as instrumental variables makes a connection with randomization, since these are generally independent of many quantities that we might investigate. The difference between living in a high and a low tax area is somewhat like that between people made to smoke purely by experimental fiat. Neither is taken to be influenced by factors that normally lead people to smoke. A random element is introduced by comparing statistics for people whose postal code ended in an odd digit and who smoke, with those for people whose postal code ended in an even digit and who do not smoke. There is also a resemblance to the potential outcomes approach: with a

107 Sussman and Hayward (2010). Pearl (2000) chapter 7 gives an abstract definition, but ignores what is for me an essential question, how one can know that a variable satisfies the definition, though he partially addresses the issue in his chapter 8.)

108 Leigh and Schembri (2004).

109 Humphreys, Blodgett, and Wagner (2014).

large sample we can expect that most profiles occurring in either group will also occur in the other.

In "Mendelian randomization", an interesting development of instrumental variables, the connection with experimental randomization is explicit.[110] This technique applies when there is a correlation between a particular trait and the occurrence of some disease, say. A sample of individuals with a genetic predisposition to the trait is examined. If the trait is a cause of the disease then it will occur more often in individuals with the predisposition than in the general population. If not, the correlation runs in the opposite direction or is due to a common cause. It is essential here that genes coding for different traits are independent of one another, so that the proportion of genetic types is otherwise randomly distributed.

Instrumental variables are a clever idea and in theory the technique can do what it claims to, if the assumptions about causal independence and about the variety of the sampled population are correct. But in particular cases it can be contested whether the variables in question are really as causally inert as required. One sign that they might not be is that investigations of the same issue using different instrumental variables often come up with rather different answers. The claim that something is a suitable instrumental variable, or that it has the independence needed in a substitute for randomization, must itself be justified, and while ideally this would be done by experiment it is usually based on prior theory or what seems evident.[111] In some versions of the technique, such as Mendelian randomization, there is a profound dependence on

110 Davey Smith and Ebrahim 2003.

111 Rosenzweig and Wolpin (2000) is an influential analysis expressing scepticism about the causal inertness of some uses of instrumental variables in economics.

prior theory. This can produce a greater danger of circularity than in experiment strictly construed. It can also produce very strong evidence, if the prior theory is assumed. So there is a real possibility of strong evidence that is not robust[112].

Another technique is "natural experiments". In biology a lot of attention is paid to evolution on islands and in isolated lakes, as substitute for experiments isolating a species from competition, identifying selective pressures on it, and then observing how it evolves ("bottle experiments").[113] A contrived experiment with the evolution of a multicellular organism would require enormous capacity to control all the variables, and often would take thousands of years. Darwin was very fortunate in coming across the Galapagos Islands.

Natural experiments along these lines are essential in biology, especially in evolutionary biology. Determining that a situation is a natural experiment, that particular influences operate and other ones do not, requires prior theory. It can be very unproblematic. It is hardly dangerous to assume that few land animals can swim across thousands of kilometres of ocean. (Though it has been suggested that snails can travel long distances in the stomachs of birds.)

Sometimes natural and contrived experiments are combined. Biologists have taken the contents of lakes that have been isolated for millennia and transferred them to artificial

112 A "futuristic" development of the technique could give *stronger* evidence than simple experiment, with the danger of less robustness. It would ensure that all members of the two groups were genetically identical except for genes affecting the trait in question. With traditional randomization the traits are on average even over all groups but can vary wildly from sample to sample.

113 Schluter (2000), chapter 3, Fraser and Keddy (1997).

ponds in their laboratories where they can be studied under controlled conditions. Cleverly discovered and artfully created controls can be continuous.[114]

Natural experiments and instrumental variables rely on assumptions of causal independence. The experiment does not reveal what it claims to unless the ersatz intervention really is parallel to what you would find in an ideally constructed experiment. These assumptions have to come from fundamental theory or from other experiments. The evidence is more solid if the basis is experimental. Otherwise, a mistaken causal assumption may be perpetuated and spread wider. Sometimes fundamental theory can make the link. In the next section I discuss how basic theory substitutes for experiment in astronomy.

Besides these failings of self-sufficiency there are also limitations. We have to take the experiment-substitutes as we find them. We cannot tune them to fit the hypotheses we are testing, for example by giving individual human subjects precise degrees of smoke and checking for susceptibility to lung cancer and heart disease.[115]

astronomy

Non-experimental evidence has its successes, and none greater than astronomy, where we have accounts of many processes which stand up well to later investigation and where experiment is generally impossible. We cannot move stars. When we change astronomical theories it is usually to refine or enlarge them. (Cosmological theories might

114 Kawecki 2012.
115 Fine-grained descriptions of causal relata are a feature of Lewis's later theory (Lewis 2000).

be exceptions.) Astronomy might seem to be an important counterexample to the special role of experiment.

I think astronomy is a special case. We should not expect it to generalize and we cannot be sure that the happy situation will continue.

the simplicity and regularity of the data The locations of the stars as seen from the earth are fixed; they do not change within a human lifespan. The motions of the planets ("wandering stars" one way or another in European languages, and for that matter in Mandarin) are more complicated, but not hard to describe. (It is *explaining* the regularities of the planets that is harder.) The movements of comets are yet more intractable. Luckily the details of these were largely ignored until astronomy had gained confirmation and power.[116]

gravity The movements of the planets are predictable, on a heliocentric model, by one force alone, gravity. Moreover the stable pattern is predictable to a good first approximation simply by considering two-body gravitational interactions between a planet and the sun. The first gravitational account of the solar system was Newton's, and his methodology involves starting with planet/sun forces and then adding more complicated interactions only as needed.[117] So we can make good and generally irreversible progress by the technique of advancing through n-body gravitational interactions, increasing n only as required.

---

116 Fantasy: a metaphysical/political situation in which for astrological reasons the trajectories of comets have to be predicted in detail, leading to the failure and abandonment of early astronomy.

117 Smith (2002), (2014).

the universality of the physics Developments in physics, testable by experiments on earth, can be applied to the wider universe. Early physics and astronomy scaffold one another, testing and lending support and prestige to one another. Since the effects of gravity can be tested on earth, for example by the famous Cavendish experiment, the stage is set for assumptions of continuity between terrestrial and universal forces. The same is true of optics.[118] Aristotle suggested the opposite: that sublunary and celestial physics are different, and the realization that this was wrong gave a tremendous impetus and freedom to early astronomy. We apply experimentally testable theories about nuclear fusion to the composition and evolution of stars.

Will astronomy continue to be so lucky? We cannot tell what developments in testable fundamental physics there will be, or what phenomena will remain unexplained in spite of all our efforts. But there are issues that have eluded us for a long time. The origin of the moon, the reasons for the spacings of the orbits of the planets, the origins of comets: these are all questions about the solar system which for which hypotheses have long gone back and forth and only slowly stabilized, as they do in non-experimental science. As for causal questions about the wider universe: we can now apply fundamental physics to many of them, and these will at any rate reduce the number of sustainable hypotheses. But many such questions demand fusions of general relativity and quantum mechanics that are delicate and uncertain. Perhaps the dark matter/dark energy conundrum will resist our best efforts. Perhaps there are irreducible forces and processes that occur only at astronomical scales, where enormous energy is available, that could

118 Some non-optical astronomical instruments draw on much more recondite theory. See Shapere (1982) for the case of neutrinos in studies of the workings of stars. There are hi-tech ways of measuring anything: for gravity and thus the weight of the earth see Xu (2019)..

only be manipulatively tested with experiments that are unlikely to be available to our species.[119]

The obvious comparison is with economics. Economics has played with a series of candidates for the role of physics,to give a background that makes experiment less necessary. The theory of rational choice and game theory were early suggestions. Their assumptions of perfect rationality proved worrying, and the emphasis now is more on experimental game theory, discovering how people actually behave in strategic situations.[120] Prospect theory, adapting rational choice to human limitations, was once very appealing but seems to have lost its lustre. Econometrics has lost a lot of the prestige that it once had, largely because it is lack of controlled experiments is not compensated for by accurate predictions.[121] There is a vogue at the moment for neuro-economics, basing predictions of choice on studies of the brain. It is too early to tell whether it can fill the chosen role. But it is far from automatic that gives a substitute for experiment in economics itself. The situation of economics is hardly a dismal prospect for astronomy, but it would be a step down from its throne.

The conclusions from this part match those of the previous two. There are costs to substituting other means for deliberate experiment to confirm hypotheses about causation. They are paid in loss of causal depth and in flexibility, the capacity to focus on a specific aspect of the situation being investigated. The loss of flexibility is more

119 Longair (2006)
120 Guala (2005)
121 Econometrics was a source of many ideas about testing causal hypotheses, such as instrumental variables, and is generally an area for great statistical sophistication. It is still an important source of techniques for handling very large amounts of data.

worrying the more that one takes causal concepts to vary in their connections with the depth of underlying law and closely related to this the specificity of the contexts they are applied to. And the greater the variety of the objective relations that masquerade under a single everyday causal concept one takes there to be, the more important it will be that there is no simple and uniform technique, such as those I have discussed aspire to being or randomization as a mechanically effective element of experimentation, to force the data to unambiguous conclusions.

chapter 7: distributed knowledge

Knowing is a form of doing. *We* accomplish a lot of knowledge, together with everything else we do. And we do much of it together in complex shared and interacting ways. This is not to deny that there is also an infinite amount that we do not know. In many domains — agriculture, building, science — knowledge involves shared and distributed activity. (So you typically cannot say which individuals in a group did something; they did it together.) The result of shared inquiry is a collective connection with the facts, and it is inevitable that experiment, the clearest example of inquiring interaction with things, has a central, if not exclusive, place in this.

The aim of this chapter is to state the attitude loosely described in the previous paragraph more carefully. This will require saying enough about the teamwork involved in carrying out and evaluating the results of experiments to show how it leads to knowledge possessed by communities of scientists. Along the way there will be more "conceptual" conclusions. Some of them concern why these matters are best described in terms of knowledge rather than rational belief. Knowledge is intrinsically suited for sharing.

two kinds of cooperative inquiry

In inquiry as in other shared activities there are two patterns. At one extreme is simple distribution of time. Coworkers could do each other's tasks and have access to each other's information. Two police officers who alternate playing good cop and bad cop, two violinists in a quartet who change roles every other performance, you climbing the ladder

while I hold it although for the next job it will be you holding it while I climb. At the other extreme is distribution of knowledge and understanding. The architect designing a clinic for the doctor needs to know design constraints that only the doctor can tell her, while the doctor has to to compromise with structural constraints that only the architect can tell him. Between these there are many ways of spreading tasks and information around .

Both extremes are found with cooperative inquiry. At the division of time end one person looks into the microscope and dictates notes that the other writes down, and they are both equally trained so that they can swap jobs when they get tired. At the division of expertise end a mathematician and a field biologist work together, neither able to understand in detail how the other does what they do. Or a team of 200 people construct, fine-tune, and operate a particle collider; the team contains theoreticians who understand in detail the eventual aim and the unexpected information that might be valuable, experimentalists who have the knack of teasing the information out of subatomic processes, engineers who can design it, electricians and others who can put it together, and the people who construct honest but intelligible grant applications. None of these tribes knows everything that allows a member of another to do their job, and all of them have the merest amateur ignorance of some of the other expertise. Yet it all hangs together to serve the collective enterprise, which only a few of them can formulate in accurate detail.[122]

Suppose the team gets an important result. Whose result is it? Suppose that it is true, and their procedure is not only impeccable in terms of the current standards but

122 Gallison (1987), (1997),Randall (2011).

appropriate in a way that an even fuller knowledge of the facts would endorse. Then, if ever, the result is knowledge. But whose knowledge? In its most complete form no one of them can even state it. *They* put together the knowledge and in that sense they know it, collectively. (And future people may say "already in the twenty-first century they knew that…", with an even vaguer and wider "they".) But how can this be when no one of them can even state it? Take a step back for a better look.

what knowledge requires

Collective or distributed knowledge is a less problematic idea than collective or distributed belief. Compare with action. Collective intention is trickier than collective accomplishment, because the latter needs only that all the components of what is accomplished be produced, in the right arrangement, as a result of the activities of individuals in the collective. Similarly, distributed knowledge needs only that the thoughts of the knowers be sensitive enough to the known situation. A set of sufficient conditions is that (a) there is a correlation between ways that the situation could have been and ways that collective states of the knowers-together would consequently be, (b) some thinking or activity of each of the knowers is essential to (a), and (c) a significant cause of each knower's activities or thinking included in (b) is the activities or thinking of other knowers, in a way that links every one of them to everyone else. There is more than a little vagueness in these conditions: correlation comes in degrees with various causes, thinking and activity are not particularly sharp concepts, the linkages of each to each can be distant or intimate. But this has advantages; it makes it easier to discuss evidence, as we will see. It is not required that the whole fact be represented in any person's mind. The distributed partial representations can then be the basis of coordinated activity on

the part of the individuals to result in a collective accomplishment, which frames further collective action, and so on as in the purely individual case.

Individual representations are often of particular things, properties, and processes, while the resulting shared representation can be of a fact, something with logical form and potentially expressible in language so that it can be true or false. And that propositional thing is the object of knowledge. For a clue about how this transformation takes place consider a case of imperfect testimony. An observer with good eyesight but not much knowledge of birds is reporting to a myopic twitcher who is taking notes. Both of them are sending their data to an evolutionary biologist who has enormous ornithological and other knowledge but is hopeless in the field. The observer says that a large dark-coloured duck with a white necklace is having trouble getting airborne from the lake. The twitcher knows that the bird must be a loon rather than a duck. The biologist knows that loons are not closely related to ducks, that they often have trouble taking off, that this is south of the normal range of loons, and that for its subspecies this bird is behaving normally. So she sends a note to a professional newsletter saying "*gavia arctica* may be extending their range southward while behaving otherwise normally". All three participants are essential, because of the limitations of the observer's skill in recognition, the twitter's eyesight, and the biologist's field skills. None of them has the full content, because only one knows exactly where on the lake the loon was, only one can identify it as a loon, and only one knows the taxonomy and life habits of loons.

The three-person team is an instance of a pattern typical of distributed cognition. One agent identifies and names the object — that bird on that lake — another then applies

appropriate predicates to it — loon — and a third puts these together in a relevant logical form. All three — name, predicate, form — will often trace back to their earlier uses in other reference-defining contexts.

The same general pattern occurs in more sophisticated cases. Someone can refer to an object of their thought, though mis-describing it, and pass the reference on to someone else, who may also be the recipient of information from others, who can put it all together in ways that correct their mis-descriptions. That much is clear from the classic causal account of reference.[123] To extend this, consider that the recipient may have an inaccurate description in some respect, and that one of their sources may have a more accurate one in that particular respect. The biologist may assume that the loon did eventually get airborne, because they usually do, while the observer saw that it did not. In these cases the knowledge is not simply shared but distributed, a different compound of components that can exist although no single participant has all of them. In the examples a dose of communication and persuasion might make the distributed knowledge available to all parties. But in more complex cases individual limitations in skill, background, and capacity to manage complexity will make this impossible. So the knowledge is distributed if it exists at all.

Scattered representations may well fit together to make false or insane "beliefs" also. But the combinations that count as knowledge are special. They should track both the possible alternatives to the facts that make them true — the more of them and the more possible the better — and the future true constellations of states that arise from similar situations. The general pattern is that states bearing components of reference to aspects

123 Kripke (1980), Putnam (1973), Burge (1988).

of the putatively known facts covary counterfactually with those facts.[124] If things had been somewhat different correspondingly different contents should have been assembled, and when things are different in the future parallel cognition, operating in and between the relevant people, should frequently lead to equally successful outcomes. (If there had not been a dark water-bird on the lake, the eventual report would not have been that there was a duck there, if it is a simplified Nozickian account. And if a duck or a loon does appear it will be reported as such.)

This is very close to a now-standard tracking conception of knowledge, as I discuss it below in the shared/attributed case. And it is vague enough that it can be reconciled with several versions of that.[125] It does not rely on a prior notion of belief: it makes its own. Later in this chapter I discuss how it can work with no conception of belief, narrowly construed, at all. And in the next and final chapter I discuss the significance for the role of experimental evidence.

series of tests by teams of researchers Scientists nearly always work in teams, generalizing over a vast range of styles and organizations. This is particularly true of experimenters, as sociologists of science have explored in detail. The teams are often very large, sometimes enormous, with complex social structure, bringing together the work of people with interlocking but generally distinct expertise.[126] Research teams

124 Speaking of facts is not meant to introduce an ontology of them. Rephrasing in terms of individuals and properties would say the same, but at greater length.

125 The grandparents of these positions are Robert Nozick and Ernest Sosa. Nozick (1981), part 3, Sosa (1999). Note that Nozick uses a somewhat non-standard conditional, and the analysis is nearer to a biconditional than would appear from this and many expositions. See also Greco (2016), Pritchard (2018), Rabinowitz (internet).

126 Gallison 1987, chapter 6

contain many specialists: equipment designers, statisticians, theoreticians, organizers, and others.[127] Missing from this list is "evidence assessor". No one is trained or functions to tell their colleagues when they should believe a hypothesis. Team leaders make announcements on behalf of the group, and before doing this take advice particularly from their tame statisticians. And in general individual researchers can get by with a very superficial grasp of what others in the team understand. As in many cooperative activities, what individuals believe about the facts can be less important than what they believe about what other people believe. There may even be cases where most people in a discipline think that most others subscribe to some view and tune their cooperation accordingly, though in fact most have private doubts. (Candidates might be attitudes to ways of papering over the cracks in standard approaches to the mysteries of quantum mechanics, or working linguists' attitudes to the latest MIT syntactical orthodoxy.)

On the other hand there usually is a consensus about which tests a hypothesis has passed or failed, with what margins, and perhaps less confidently how stringent they were. The criteria for these are hammered out before the experiment is actually performed. (Statisticians typically come in here, setting up a situation which will produce results which can be handled in a prearranged way.) There is also usually a consensus about the theoretical advantages and problems of a hypothesis. But we need much less of a consensus about how all these tests and advantages are to be weighed against one another. Then, well into the process, the team can decide such things as whether to publish, what to say to other groups, whether a press release is called for, and so on. To do this they do not have to agree about how to sum up the force of evidence, let alone

127 Perhaps the first research assistant was Alexander sending samples to Aristotle. He wouldn't like that description. Neither of them would.

whether they have established or refuted one of their hypotheses. What they have to decide is whether an action such as publishing would be good for the subject, good for them collectively, and good for them individually.

In a typical history one team of researchers tests a hypothesis experimentally, and then another team replicates the test, or tests it in some other way. Eventually there is a series of tests of varying stringency which the hypothesis has passed or failed to varying degrees. After all this there may be a consensus among scientists: established, refuted, still up in the air. This process can be a charade, if the tests are badly done or if there is some mismatch between them, as when they mistakenly use different medicines, species, substances or whatever, for example when tests of the medical effects of compounds ignore the difference between isomers. The probabilities may also be inappropriate. These issues were discussed in chapter 4.

Different investigators and different teams of investigators will react differently to single tests or small sequences of tests.[128] Different theoretical backgrounds and different commitments to what the satisfactory theory should be like will result in different assessments of the evidence. Sometimes what one school takes as finally establishing an idea will be taken by another as a target for refutation and undermining. Sometimes this can mean resisting the force of evidence, and the constructive aspect is searching for experimental evidence and theoretical constructs that allow the reassessment of the evidence so far. It is a bad idea to give up too early, just as it is a bad idea to be too stubborn, to give up too late. As Kitcher has argued, it is a delicate business whether one

128 While agreeing about which tests were passed and how severe they were. This is like the distinction between evidence and confirmation in Bandyopadhyay, Brittan and Taper (2011) but with more cynicism about confirmation.

should be responsive or stubborn in the face of evidence, for example a series of varied tests, that could still be resisted. Among other things it depends on the structure of rival positions in the discipline, how equally populated they are, whether there is a consensus that can be altered, and how subject to reinterpretation disputed evidence really is.[129] I should add to what Kitcher says that the value of one or the other strategy is not determined by what individuals think about the force of evidence and the structure of the discipline but by what the facts about these things are. Retrospectively, perhaps years or centuries later, historians of science may be able to say "that was a good move" or " that turned out to be a mistake".

When the process has gone well apparent and real stringency are not greatly separated. There is a fit between the results of particular tests and the causes of the phenomena that feed them and as a result there is a fit between successive tests. They have the same targets. When all goes well, as a result of correctly applied standard procedure, the results of different tests carried out by different teams refer to the same phenomena and their causes. Then no one has to sum up the series in the form of a verdict, and the shared objects of reference can be part of the basis for shared or distributed knowledge. In the ideal case, the two dimensions of fitting, between test and fact and between test and further test, underwrite a knowledge-based reference: the terms in the hypothesis refer to corresponding qualities of objects, which qualifies as shared knowledge of them.

meta-analysis

129 Kitcher (1990), Wray (2002, Perovic (2011).

There is an instructive analogy between shared knowledge based on a sequence of tests and the somewhat controversial statistical technique of meta-analysis. In meta-analysis we take a number of studies of the same hypothesis and combine them in ways that promise to give more reliable results than any of the individual tests. This section is not the in-depth analysis of meta-analysis that waits for the attention of some statistically sophisticated open-minded philosopher of science. Instead the question is more limited: what we can learn from meta-analysis about combining studies.[130]

In the very simplest kind of meta-analysis one simply combines the data from the component tests and performs the same kind of analysis to this larger body of data. Even this procedure can give striking results. The result can be different from the majority, or even all, of the component studies. It is not just an averaging of their outcomes. The reason is that it is based on a much larger sample than any individual study.

We can make this crude method more sophisticated by building in a comparison of the sizes of the effects of the relevant variables predicted by the hypothesis with those estimated from the data, and the variances of the effects about their means in the particular studies. Then we can average the effect sizes and combine them while giving greater weight to studies with less varied outcomes. The result is usually called the fixed effects model. We can increase the sophistication further by including a measure of how much the samples vary from one another. This can take account of the possibility that the samples in the different studies were drawn from somewhat different populations. A further layer can incorporate judgements about the methodological quality of the various studies.

130 Hans (1997), Hunter and Schmidt (2004), le Lorier and others (1997), Stegenga (2018) chapter 6.

Meta-analysis transparently requires the results of a number of studies, whose details are available to a corresponding number of teams, to be combined so that a compilation of their results is available to the person or team doing the meta-analysis. They are unlikely to work with the full details, and in fact it is likely to be counterproductive if they do, though it is important that they can access them if required. And the results of the meta-analysis will be shared with investigators who have not the sketchiest grasp of the individual studies. So even if there is only one-meta-analyst, the full contents of the studies is shared between that person and various consumers of the result

These consumers will often not have knowledge of the result — as individuals — because of these problems of transmission. But in many such cases there will be knowledge spread between them and the original research teams. For the full facts about the individual studies and their meta-analytic glomming together will be distributed between all these minds. Eventually this can take the form of knowledge possessed by individuals. It certainly will not always.

the epistemic irrelevance of individual belief

What communities think and what is handed down to later generations is more important than what occupies the mind of any particular person. This is true in science and in unscientific matters. Focussing on the process leading to general acceptance we can distinguish four stages. Belief is not needed at any of them.

The first is motivation. A question arises from previous research or from a worry about existing theory or from evidence from tests of a hypothesis that are inconclusive or unsatisfactory in some other way. So more experimental work is needed. Experimenters may not have the question, worry, or lack of satisfaction because they believe the hypothesis is false. It is enough that they think that others could be unconvinced of it. (And there is the point already made that what everybody thinks everybody else thinks is more important than what they really think.) So there is a motive for planning experiments, which may involve considerable thinking and ingenuity. Secure belief in existing explanations can be involved here, in thinking that a proposed experiment has a good chance of revealing something. But it is not belief in the hypothesis concerned, or for that matter belief that it is false.

The next stage is to carry out the experiment. Belief or disbelief is obviously not central here, since the issue is whether. Equally obviously, thoroughly established or refuted or nutty possibilities are far from any priority. (There is no shortage of interesting live conjectures.) As with the planning stage, shared established doctrine may play a large role in doing the experiment, but this is not belief in the hypothesis or its negation.

The third stage is analyzing the results. I have been describing this in terms of passing and failing tests. One could extend this to include gradations of passing or failing: strongly, marginally, and so on. These are not degrees of belief! A hypothesis can pass a test strongly while its negation passes a different test equally strongly. Confusing but possible.

Likelihoods may play a large role in analyzing the results. These also are not degrees of belief. They can be intrinsically connected to the hypotheses, saying how likely the data is *according to* the hypothesis. Or they can be an agreed consensus among the relevant scientists, more of a convention than an average of what they will think. As James Hawthorne says

> Students are trained up on examples that instill a capacity to correctly perceive the implications. This training ultimately tends to provide a high degree of expert agreement regarding what a theory says about specific cases — that is, a high degree of expert agreement on the values of the likelihoods, or at least on values of likelihood ratios. These agreed likelihoods do not simply represent an expert's individual degree-of-belief that an event has or will occur if a given theory is true—at least not in the first instance. Rather, it seems, experts consult the likelihoods they have internalized through their training to help them determine what to believe in specific cases, and how strongly to believe it.[131]

"what to believe" here means what likelihoods to assign. It can differ from the strength of a person's expectation, because they depend entirely on the content of the hypothesis. Other sources of information or opinion are disregarded. As he goes on to say, otherwise

> ... each member of the scientific would generally have his or her own distinct personal likelihood for the evidence, depending on what else he or she knows about the evidence. This would make a complete hash of scientific hypothesis testing. What should a researcher report in a scientific journal?

The aim is to produce results that other researchers, perhaps from contrary traditions, can use to guide further work. And

131 Hawthorne (2005).

> Bayes’s Theorem is central to Bayesians precisely because they treat the likelihoods as more objective than the posterior probabilities that likelihoods are used to calculate. Bayesians treat the likelihoods as stable points to which prior and posterior probabilities must coherently conform.

The use of ratios of likelihoods in "objective" Bayesian methods is no problem, since they are obtained by procedures that others can replicate rather than coming from the vagaries of individual minds. (Prior probabilities for hypotheses and even for data are not problematic either, as long as they are obtained in simple ways and they are thought of as devices for getting the likelihoods right.) What we do not need are beliefs that vary from person to person although these people are engaged in deeply cooperative enterprises.

In the fourth stage hypotheses that have done well in series of tests are presented to the larger scientific community and to the whole intelligent world. This generates beliefs, propositional states that would be knowledge if their content was adequately connected and supported. Often they do not qualify as knowledge, because the conditions are not met. (A student hears a marketing class by a thoroughly expert speaker, but is not really paying attention though he is lucky enough to absorb this claim, which he could easily have garbled, and has no sense of why it is true.) The agents at this stage are textbook writers, popularizers, survey article writers, and so on. They are usually reflecting a consensus that a body of scientists has arrived at. But they will very rarely know about all the experimental work behind the idea that they are disseminating and all its presuppositions. So taken in isolation these people also do not know. Taken as representatives of communities they have a form of shared knowledge.

When the path to orthodoxy takes these four stages, at any rate, we have a disjunction. The item is either known in an essentially shared or sometimes distributed way, or it is not knowledge at all.

Consider a hypothesis that receives the imprimatur of the varied tribe of codifiers as a result of passing a series of stringent tests, and eventually filters down to graduate students and historians of science. When enough of this has happened the world thinks of it as having been accepted. But the acceptance may be premature. The trivial reason is haste or enthusiasm, possibly encouraged by an attractive hypothesis promising explanatory power. A deeper reason is tests that are less stringent than they seem, because of errors about the background conditions of the experiments, or because of unknown factors influencing their results. The opposite can also happen. A hypothesis can get to this stage in the mind of the consensus-making tribe although in fact the tests are *more* stringent than they are thought to be. In either case, the point in time when the hypothesis is generally accepted can be different from the point when there is strong accumulated evidence for it. So much the worse for the idea that there is a fixed level of evidential support beyond which it is wrong not to accept a hypothesis.

degrees of knowledge, degrees of sharing

The ignorance/knowledge contrast and the shared/distributed contrast are not as absolute as it may have seemed. In both science and common sense we have somewhat flexible standards of both, and the standards for each are connected. I believe this is

generally accepted, when expressed right. So the task is to say it clearly enough that the obviousness becomes obvious.

Professor Shirvani, a world expert on gravitational waves is giving a lecture on which aspects of general relativity are involved in detecting them. She makes a controversial and somewhat technical point, with a defense of it that would have convinced any of her expert colleagues. An undergraduate physics major is listening to her. He either

a) understands and accepts the claim, but was not really paying attention and could easily have misunderstood it.

or

b) is trying hard and succeeds in understanding the claim and its justification, but the exposition is so technical that a misapprehension was only avoided by chance

or

c) is paying attention and has the right background but Shirvani is a terrible communicator and it is only by luck that the correct sense gets through.

Does he come to acquire or share in knowledge? I think it is intuitively clear (and fits with the analysis that follows) that a) is an extremely marginal case of knowledge on any standard conception and that b) and c) qualify as knowledge or not depending on where the boundaries for it are drawn with a generally admissible range. Within this range, moreover, once it is admitted that the case is somewhere within it, the knowledge/ignorance conclusion depends essentially on where equally acceptable boundaries are placed.

It is a complicated vagueness, though, depending on a number of factors. Three of them are:

– the tightness with which the resulting state tracks the fact through possible ways it can occur and can fail to occur

– the reliability of the chain of transmission from one user of a word to another

– the tightness of the connection between knowledge in a particular case and true belief (or knowledge) in similar cases.

The first, tracking, measures the because-ness of the state, how much it is true as a causal effect of the fact that it is true.

The second, transmission, measures the unity of the thinking spread between different people.

The third, similar cases, is a central reason why knowledge is an important concept. It underwrites the connection between knowledge and evidence.

While the knowledge/evidence connection is a large part of the importance of knowledge it does not help decide where the boundaries should be drawn. For there can be evidence, even strong evidence, for hypotheses that fall far short of knowledge. A theme of this book, returns: characteristics of knowledge are found even when knowledge is not attained, notably in the relation between evidence and hypothesis.

Begin with tracking. The representation that is a candidate for knowledge should correlate with the not too exotic possibilities for its truth. The correlation does not have to be perfect, but there has to be a causally substantial link between facts and their representations. Not too many exceptions, and preferably obtained by isolated unsystematic processes. The allowed exceptions vary from one modal epistemology to another; I am bundling them altogether to be managed with the auspiciousness considerations below.

When dealing with a mythical isolated agent the tendency is to postulate a single representational state, a belief, which incorporates the entire known representation. But this is not necessary. Consider the three-person loon-watching example again. The biologist does not have to know the details of what the ornithologist and the observer report to her. She only needs enough that she can put together an outline view that can serve in consultation with them. If her summary account is in touch with the loony fact, then were the observed fact different the observation and the tweaking of it by the ornithologist would be different and so her outline would be different in its full content. In fact, it would be enough if she simply reported that a bird as characterized by the twitcher was in the location specified by the observer doing what the observer said and thus leading to the conclusion drawn by the ornithologist. So the basic requirement is just that among the interacting states of mind of the person or people concerned there be states that have the right responsiveness to possible variations in the actual situation. For different facts they can be states of different people, and for any given fact they could be different kinds of state of any given person.[132]

132 So we do not need belief as a single category of states of even isolated people.

The effect is to make the correlation conditions bear the weight otherwise supported by the concept of belief. This satisfies a somewhat arcane and definitely controversial need in the case of a single knower, but is essential in the case of a network of cooperating investigators.

Now auspiciousness. We want to locate the exceptions to perfect causal correlation between representation and represented so that present knowledge tends to future truth. No real human being exhibits a perfect correlation (probably not even with any single conclusion), but the demands of different topics and modes of thinking are different so that the concessions to human imperfection are different for different kinds of knowledge. They are likely to be different for quantum mechanics and field biology. The need is partially met in terms of the correlation requirement left vague just now. The allowable exceptions to perfect correlation between ways that a hypothesis could be true and ways that a person or network of people could arrive at it as a conclusion, in order for that conclusion to count as knowledge, can be stipulated so that in similar situations the same means also lead to true conclusions. That is, we can declare that the right set of exceptions in order for a particular hypothesis to be known by a particular person are those that maximize the chance that in similar circumstances any conclusion she comes to will be true.

At least some of the additional burden is taken care of by the fact that all such loosenings of the correlation only make exceptions to an overall correlation. So whatever form they take knowledge will still generally result in true belief when circumstances change. This obviously makes weaker “knowledge” still worth having. Its value remains when it is

scaled-down: weaker tendencies to true belief are still desirable for researchers when that is the best they can get. Even the very beginning of the road to is still being on the right path.

Possibility c) above introduces another matter of degree. The transmissions of reference and of evidential force from one person to another may be more or less strong. When what is passed on is extremely weak there is very little to distinguish as distributed rather than individual knowledge. And they never get as strong between people as they do within one person's brain. This is another consideration tending to the conclusion that there is no absolute fixed point at which general contact with the facts turns into knowledge. But, as I have been emphasizing, this makes my larger position easier to argue for, since it opens up a slippery slope down to degrees of "knowledge" where we do not apply the term, where thinking in terms of evidence rather than any form of knowledge is more natural. People can be on the road to what we do call knowledge even if they are not far enough along for complacency about whether we have arrived.

Scientists and everyone else

Science is one of the main homes of distributed knowledge, in part because the activities of experimentneed and it are well suited for cooperation and shared intention. Though I think there are analogues in pre-scientific life I will not argue for them. Instead I will note how underappreciation of the collective element in science can complicate an outsider's attitude to scientific results.

An intelligent non-scientist with a decent but entirely non-participatory knowledge of science reads of a new results. It could be a newly congealed synthesis which has been put together and is generally accepted, or it could be a test which a hypothesis has passed or failed. She is likely not to understand where the idea is placed in the cycle from conjecture to acceptance. A crude consequence is that she will not know whether to think "this is what they are asserting now" or "that is one piece of evidence, no doubt among many, for it".

Moreover, she will not have much idea about the strength of the evidence that any particular trial gives or what the alternatives it was tried against are. Both of these require experience of scientific practice and of the recent give and take in the area in question.

Scientists in other fields than the particular topic will often be in a similar situation with regard to what they do not know. But they will be much clearer about the fact that they do not know these things. And their experience in fields that work in similar ways will allow them to grasp and discount their ignorance of the details. Combined with this, they will have learned how much in their own work they rely on the competence of others. (This makes scientific fraud easier, since being set up for distributed action science is a generally trusting area.) So if they understand that the methodology was standard and the researchers well qualified, they will have a fair amount of confidence in the outcome. All of this may sometimes make them too accepting of overconfident work, out of ignorance about the details of what actually occurred and the people who did it, but it will not lead to blanket distrust of what they find from well-informed sources. (Think of

Nature or *Proceedings of the National Academy of Sciences*, which researchers often read to keep up with developments in other fields. Philosophers will often read such journals, but those completely outside science, however generally well informed they are, will pick up most of their acquaintance of suggestions that are well supported by evidence from sources that mix it up with epistemic junk.)

So not participating in distributed inquiry, non-scientists are vulnerable to misunderstanding what lies behind the results they hear of. This generates a general mistrust.

and experiment

Knowledge and epistemic cooperation suit one another, in that sensitivity and similar properties of knowledge link the known content to its objects independently of the vagaries of a particular person's use of language. The network of cooperating inquirers is nearer to self-sufficiency. When the cooperation involves experiment the congruence is stronger, even more so when it is a matter of distribution rather than simple sharing.

Think of a cooperative inquiry as like a radio telescope with an array of widely spaced receivers. Or like a school of individually vulnerable animals who by gathering together can pool their sensory resources. When each step of observation and reasoning is carried out by a separate individual, as with shared inquiry such as the loon example, then the chain can fail at its weakest link. (If the ornithologist takes the loon to be a coot rather than a duck then all the skill of the observer and all the expertise of the biologist are

wasted.[133]) But when the roles are distributed and overlapping the results can be checked against one another.

Parallel considerations apply to designing and performing the experimental process and analyzing the resulting data, the articulations of experiment from chapter 2. Performing the experiment is in general almost trivially distributed, in that it is an intrinsically multi-person action. Moreover in different situations, crucial to evaluating the outcome as knowledge, the interactions of different contributors, and indeed who contributes what, can be different, still leading to accurate results. Planning and analysis are much like observation. Agents can perform sub-tasks without knowing the results of one another's work. Think of a statistician outlining a number of calculations that need to be performed and delegating them separately to that many calculators who hand their results to a number of different people. And some agents can modify the tasks done by others. And if things done by different individuals conflict the tasks and the distribution can be varied.

Again the bottom line is that what happens intellectually depends on what the participants do, and that their collective doing is no some of individual results. Only then is the full power of experimentation exploited, and together with this the information gathered exceeds both what any one participant could collect and what they could collect together by simply handing along partial information from one to another.

As a byproduct, we get a sharper description of the contrast between distributed and merely shared knowledge. A conclusion is distributed to the extent that participants are

133 A coot can be a foolish person, just as a loon can be a crazy one. And a goose is sometimes someone silly, a chicken someone who is unreasonably afraid. What is it about these bird words?

influencing one another's results rather than either simply receiving them or simply adding to their effects.

ch 7: evidence, finally

Why do we think that nothing travels faster than light? Because we have such a lot of evidence, much of it derived from experiments. How do we know that nothing travels faster than light? Because we have *done* the experiments. Two entwined themes have run through these chapters. One has been the connections between experiment and desirable features of evidence. The other has been the presence of knowledge-like attributes behind the scenes of a number of epistemic phenomena. But mingled as they have been these two themes have not yet merged. How do they connect? What *is* evidence? What general characteristics of experiment bring these benefits? Now is the time to tie the threads together.

K-evidence

Evidence is the raw material that builds into knowledge. When you have enough evidence you stand to know something. (Other conditions obviously have to be met also, in particular truth.) I suspect that all accounts of knowledge and of evidence acknowledge this to some extent, but it is characteristic of the present and similar attitudes that they make it central to the concept of evidence. More specifically the accumulation of some kinds of empirical data, typically effects of some causal factor that also results in a hypothesis being chosen or accepted, puts individuals and collectivities in a state where they have knowledge of that factor.

Chapter 0 distinguished a kind of evidence which I called K-evidence since it is essentially connected to knowledge (in contrast to R-knowledge, whose most basic connection is

with reasonableness). But chapter 6 argued that what we usually call knowledge is just the visible spectrum of a wide range of knowledge-like conditions, which range from the easily attained to the very demanding. More demanding conditions then are needed by every day knowledge-attribution are often met when data is collected in the context of an experiment. Actually occurring data d is K-evidence for H_1 rather than H_2 when:

- there is a world w_1 such that w_1 is the nearest (H_1 & d) world & for *all* w_2 if w_2 is the nearest (H_2 & d) world then w_1 is nearer than w.

and

all worlds *relevantly near* to w_1 are H_1 and all worlds relevantly near any w_2 are H_2

and

in *appropriately many* H_2 worlds H_1 is believed and in *appropriately many* H_1 worlds H_2 is chosen. (Choosing a hypothesis can be a variety of actions from taking it seriously to writing it in textbooks and teaching it to your children, depending on the stage of evaluation, as discussed in chapters 3-6.)

The deliberately vague terms *all*, *relevantly near* and *appropriately many* are to be understood so that stronger relations of knowledge require nearer and more instances, and so that the overall conception is directed to safety (that is, so that when Hi is chosen it is often true).[134] Note three basic points.

There can be evidence for false hypotheses. This may address a worry about objective accounts of evidence. There are two reasons. The first is that the message of an experiment, in terms of K-knowledge, is, or better suggests, not that a hypothesis is true

134 This is more suitable for empirical evidence, where much of what is supported is in various ways causally necessary (Ch. 5).

but that one hypothesis is less distantly actualized than another. This is perfectly consistent with their both being false. One way that this can happen is when two hypotheses are being compared and the truth consists in a third one.

The second reason is more subtle. It is that the results of experiments never simply dictate conclusions. Otherwise their familiar variation would be paradoxical. Instead of giving experimenters opinions they suggest or encourage them. They are more like the testimony of not completely reliable witnesses, diverging from one another way, than like unimpeachable authorities, even authorities giving nuanced guidance in terms of their degrees of confidence. So when a hypothesis becomes the front runner after one or even a series of experiments experimenters should still take it as giving something tentative. For the conditionals that underwrite the interpretation of the experiment are themselves tentative, and can be discredited by conflict with other outcomes. So evidence as understood here is a matter of shaping the outcome of a particular test — which hypothesis passes and how strongly — rather than how near the favourite one is to acceptance.

The vagueness of the italicized terms can be an advantage. It allows them to be applied with different degrees to different stages of the acceptance of a hypothesis, as described in chapters 4 and 6. And it matches the uncertainty of most ascriptions of evidential force.

The evidence relation is constructed here around a competition between two hypotheses. So it describes comparative rather than absolute confirmation. One advantage is a

connection with the literature in statistics. Another is the continuity between sensitivity-like and safety-like interpretations of knowledge, in replacing not-$H_{1(2)}$" with "$H_{2(1)}$". A further advantage is given by the considerations in the next section.

connections with experiment

The structure of experiments produces a reversal condition (as described in chapter 2), which produces K-evidence. The argument for this inevitably makes a few simplifying assumptions.

To see when experiments lead to reversal conditions suppose first that we have an experiment with a trigger t and two incompatible targets T_1 or T_2 linked to two incompatible hypotheses H_1 and H_2 is true, so that "if (H_1 or H_2) then (if (t and Hi) then Ti)" is true, for i = 1, 2 (the trigger will lead to a corresponding target when one of the two hypotheses holds) and if Ti occurs then Hi is chosen. Assume also that the choice of a hypothesis depends only on the result of the experiment, and that the targets Ti are mutually inconsistent, not just given the trigger so it is not the case that T_1 and T). Now suppose that t is activated and T_1 is the result. Under the supposition that one of the hypotheses is correct, since T_2 would have resulted if H_2 were true and the T_i are incompatible, H_2 is false and the true one must be H_1. So "if H_1 or H_2 then H_1" is true; H_1 is closer to actuality than H_2 is.

The argument can be applied to any n hypotheses also. It reverses conditions where the truth of one hypothesis on a finite list results in a corresponding target, to the conclusion

that if a target occurs it must be as a result of the corresponding hypothesis. This is consistent with the possibility that neither hypothesis is true, and with the possibility that the particular target observed is a result of the truth of neither but something completely different. What we have is evidence for the winning hypothesis, rather than even degrees of proof. And these considerations by themselves do not tell us how widely spaced or how near to being actual the hypotheses are. If the experiment is comparing three or more hypotheses they may not be evenly spaced.[135],[136]

Now to go from a reversal principle to K-evidence. The aim is to show that given a reversal principle for two hypotheses and an experiment favouring one over the other, the favoured hypothesis is more nearly known than the other. The central part of this is already in place, in the link between experiment and reversal. A fact is K-evidence supporting H_1 over H_2 when drawing the line between knowledge and non-knowledge between H_1 and H_2 would make H_1 known and H_2 not. That is, both may be false and the K/not-K division consists of an extension of "actual" to include situations near to actuality plus a tolerance of more exceptions between truth and acceptance than would be admissible for what is normally called knowledge. Since the standard counterfactuals I have used to describe experiment require that the consequent be true in *all* nearest worlds where the antecedent holds, there are no relevant exceptions to consider. (This is clearly an idealization.) And in the case of only two competing hypotheses there is no contrast between sensitivity-like and safety-like since on the — also idealized — assumptions in play accepting or rejecting one means rejecting or accepting the other. So

135 Choosing between more than two hypotheses might be thought of as a series of binary choices, but it might matter in what order they were taken since analogues of the intransitivities of majority voting might arise.

136 Stronger assumptions might impose more constraints here. One likely consideration would be how possible other causes of the targets are, which depends in part on the insulation of the experiment.

counting nearby worlds as honorary actuality is the only remaining factor. Then comparative nearness to actuality is all that is needed to allow the line to be drawn. As a result, subject to these assumptions, a standard experimental setup gives K-evidence for the hypothesis that it endorses.

These two connections together describe part of why experiment gives particularly good evidence. The circumstances where the experiment gives thumbs up to a hypothesis when it opposes a particular rival are ones where it is nearer to being true than the rival. With many other forms of support, on the other hand, the assurance that what is taken for evidence does in fact lead in the direction of truth for this particular hypothesis is rather indirect. It can consist in crudely inductive support for a blunt generalization about hypotheses compared with a methodology (such as one of the rival statistical paradigms, or scientific method as reconstructed by philosophers) covering a large variety of hypotheses concerning a large variety of areas. Or it can consist in an appeal to formal rationality, without much assurance that rationally false beliefs tend to be true. (The historical changes in conceptions of scientific method ought to make one pause about this.[137])

In contrast, the K-evidence approach is directed at the particular circumstances of a particular theory-accepting situation. It does have its own price, inevitably. It is hostage to unknown, and even unknowable, factors. When our understanding of the causal nature of the evidential situation or of the circumstances where the hypothesis and its rivals would be true or false is very flawed, our beliefs about what we know and what in this

137 Sometimes the appeal is to statistical procedure with the implicit defence that the conclusion is probably true, even from people who understand the difference between probability and likelihood. I hope the discussion in chapter 4 has made this line doubtful.

sense of evidence supports one hypothesis over another can be quite mistaken. The different advantages and costs of the two approaches are best resolved not by adhering to either rather than the other, but by thinking out which problems each of them addresses best and by investigating how to combine them, how to think of them as complementaries.

These considerations combine with some more scattered factors. (They are probably individually not novel or surprising but the combination may be.) The first is the fine-tuneability of experiment. The control of the trigger and the insulation in an experiment allows both deeper and wider knowledge. When data is collected under purely observational conditions there is no control over the variety of its causes, which can be very disparate. Insulation allows the experimenter to limit the range of causes (though rarely to exclude absolutely all but the factors of interest), while an artificially created trigger allows a guarantee of the presence of one particular cause. (Randomization provides a way of getting both of these.) If insulation inhibits the usual causes then the more remote ones will manifest themselves. These factors are enhanced by the opportunities of more accurate observation provided by controlled experimentation, so that causes revealed by inhibiting normally dominant factors can be detected. Then K-evidence generates N-evidence that addresses the peculiar context- and purpose-relativity of causal hypotheses.

One source of the ability to fine tune experiments is the freedom to tinker with them until they work as intended. We can try out preliminary, small-scale, or alternative versions,

varying them in an *ad hoc* manner until we get one that gives consistent results.[138] Then the results of this fine-tuning constitute the basic design for an "official" experiment. For example the persuasive evidence for the bacterial theory of gastric ulcers rests on such carefully contrived experiments rather than on the brilliant but merely suggestive work of Marshall, who infected himself in order to see the results.[139] And there is ample evidence from historians and sociologists of science that in preparing an experiment scientists extremely often vary their procedures and apparatus until they can trust their results.[140]

The other factor is the greater knowledge that experiment gives of what its ingredients are. The causal skeleton of an experiment — trigger leading to target when the array of possible causes is limited — is surely very common. And when we perform an experiment as part of an inquiry we normally know that the trigger has been activated by our own actions and that it is the cause of the target corresponding to the chosen hypothesis. We set things up to ensure this. And we usually exploit causal knowledge that is simpler and less controversial than the intended results. Without this background knowledge outcomes can still be K-knowledge, but they are not appreciated as such, so that they neither motivate nor underwrite the fruitfulness of further investigations. (Both background causal information and knowledge about one's evidence are discussed more, separately, below.)

---

138 In my interviews with experimenters at UBC I found that nearly all of them do a lot more in the way of trial runs and exploratory experiments, and put more energy into improvised improvements of their experimental setups, then is ever explicit in their publications.

139 Marshall and Warren (1984), Hsu, Lin, and Graham (2015).

140 For example see many of the cases discussed in Gallison (1987), particularly that of the discovery of cosmic rays in chapter 3, and the cases in Franklin (2016) particularly Millikan's measurement of the charge of the electron, and the Michelson-Morley experiment.

suggestiveness

Experiments often suggest further experiments. And they suggest additions to and interpretations of established theories. This can happen early in the evolution and testing of an idea; suggesting further experimentation is most useful fairly early on. But there is a complication. A person who has objective evidence may not realize that they have reasons for belief, or the strength of what they possess. And it is harder for what you are not aware of to motivate your activities.[141]

When the evidence comes from experiment there is a route from evidence to knowledge of evidence. It rests on the social (distributed) character of experiment. Experimenters have to communicate in order to cooperate, so they have to be capable of putting their observations and analyses into words. Not all of these, but enough to enable others to do their parts. When you can say something you can know what you think, and inasmuch as you are expressing it you can know what your relation to the objects of thought is. This is partly because one notices what one is saying, and partly because many of the intentions that speech-acts are conscious. As a result the status and force of the evidence becomes available both to you and to others, to the extent that you have managed to articulate it. The link from fact to brain to consciousness is fundamentally changed by the experimental set up, which within limits one can tune as desired. So it can be tuned so as to produce not just data and analysis but also awareness of data and analysis.

There is a connection here with the way that reversal principles apply to probability, as discussed in chapter 3. Probabilities there entered into the weightings of factors influencing an outcome within and outside an experiment. Those within the experiment

141 This may not always have been obvious because of my sometimes bland use of the word "evidence".

are the result of the conscious planning of the experimenters. Inasmuch as experimenters consciously know the design and workings of their experiment, making them capable of communicating it, they are aware of the constraints they have placed on the way the data is produced. So to that extent they are aware of how likely it is that the outcome will have been shaped by one influence or another. (This is certainly not total awareness of the causal situation, which is beyond human aspiration.) Thus the practicalities of experiment encourage awareness of the influences on the outcome as well as that outcome itself.

But these considerations apply only to the results of experiment, rather than to objective evidence generally. It is only with experiment that K-evidence and R-evidence come into alignment. As a result evidence has greater power to suggest additional experimental work that would be enlightening, and to influence how insertions into established theory depend on reliable processes. It hardly needs saying that these are useful in an opportunity-opening way. I think of it as basic to science and incidental to everyday belief where it is usually a refinement of inflexible innate mechanisms that are reliable only within fixed limits. (While science for all its potential power is liable to overly optimistic inference to the best explanation unless disciplined by experiment.)

bloat-avoidance

When experimental evidence it is understood as I have been suggesting we see how we can separate evidence for different parts of a complex hypothesis.[142] We often have

142 The issue is best known from Clark Glymour's (1980, Chapter 6) bootstrapping account of confirmation, which has convinced many that there is a serious problem with once-standard accounts of confirmation. It is generally agreed that Glymour's account does not work, but that something of the kind must be right.

strong evidence for a not too adventurous hypothesis and weaker evidence or none at all for a larger and more daring hypothesis that includes or entails it. We are just about certain that there were dinosaurs on earth between 200 and 66 million years ago, but much less certain that the end of the dinosaurs was caused by an asteroid strike towards the end of that period. The second is a more specific and more general claim, and it is based on more complicated evidence needing more vulnerable assumptions. But theories have trouble capturing this difference. The basic reason is that, from the inference to the best explanation down to Bayesian accounts, they focus primarily on the consequences of the hypothesis that can be collected as evidence. So when a stronger hypothesis entails a weaker one it has all the consequences of the weaker one and more. So it inherits the explanation-based evidential power of its weaker offspring.

When we think in terms of experiment we can test the weaker theory without testing the larger theory. In the dinosaur case, we can do geological and radioisotope tests on the fossils in order to date them. These do not test hypotheses about the extinction; they will show no difference between the core hypothesis augmented with different explanations of the extinction.

When there is no obvious experiment to test the weaker hypothesis without also testing the stronger one it is simply up to us to devise one. It may strain our technological capacities and our ingenuity to find ways of intervening to do this, but we can try, and setting ourselves the scientific ideal of parsimonious theorizing we will often eventually succeed. Of course we may already have K-evidence of the un-bloated hypothesis, without knowing it. But devising an experiments and understanding what we are doing

allows us conscious access to the evidence, and fine tuning/tinkering allows us to direct the experiment at the target we need for it.

non-circularity

Holistic rhetorics of theory-acceptance raise worries about circularity. A new belief is acceptable in terms of the coherence of the combinations of it and other theories were is accepted. But then the value of independent guesses may be judged by letting each support the others.[143] The use of experiment to de-bloat theories by giving relatively direct connections with the origins of the data will protect against this.

The possibilities for circularity are much less with evidence derived from experiment.

One reason is the defences against bloated theory, just described. Evidence will less often and accrue to extravagant theories, favouring particular sub-theories of them. One route to circularity is thus blocked. Moreover an experiment compares a hypothesis, or a pair of hypotheses, against a body of data directly, without referring to theory except inasmuch as it mediates the comparison.

These protections are not foolproof. The danger I shall discuss concerns statistical models. I have argued that they are usually rather general causal hypotheses. They are rarely stated very explicitly. The danger of circularity is obvious. Experiment gives a much-used protection here. (It is also protection against circularity in hypotheses generally, but I will not discuss that.)

143 Worries about this were one of the motivations for foundationalist epistemologies in the first half of the twentieth century.

The core idea is that for designing an experiment we need something less ambitious than an explanatory theory, sustained by straightforward and direct evidence. Such causal models are typically made by producing a limited number of equations most of whose variables are observable and designed to give estimates of a small number of non-observable variables suitable for the numerical needs of a particular experiment. Often variables needed for a general account of the origins of the data are omitted if they are irrelevant to the contemplated experiment. These are tested against real data before use, to give confidence in using them for these experiments. For different types of experiment different models will be needed. This is frequently possible and allows experimentation which tests more explanatory theory.[144] If the theory survives testing then it rather than the model gets the credit. The model can then be discarded, and further testing often uses different models.

Two further factors increase in the protection against question-begging circles. The first is the structure of experiments as described in chapter 2. If that is anything like correct, there is a core of an experiment that consists not of anything conceptual or theoretical but of a process occurring with a definite input and definite results. It is always possible to replicate this indicative process and reinterpret it in the light of subsequent or different theory.

144 This is something that philosophers and sociologists of science have become increasingly aware of. Early works are Hacking (1983), Franklin (1986), Gooding and others (1989.) A recent example from climate science is Steyn and Galmarini (2008), and in recent philosophy of science see Cartwright (1999), Morgan and Morrison (1999) especially chapters 1, 2,3, Wilson (2017). Alirio Rosales suggests to me the example of Einstein's prediction of the movement of the perihelion of Mercury which contrary to the way the story is sometimes told, did not involve a deduction from the full equations of general relativity but a model constructed by simplifying them under restrictive assumptions.

finding experiments

To break out of evidential cycles and split up bloated hypotheses we need to devise experiments. Sometimes they have to use novel strategies. But this is something that is always essential in science anyway. So we can welcome the need, as pushing towards a helpful frame of mind.

Meselson and Stahl's isotope approach, referred to in chapter two, could also be used as a test of the elements that are components of inheritance, separate from the whole DNA hypothesis. Cajal's application of simple staining techniques to neural tissue to make the connections of neurons visible independently of assumptions about the functions of neural networks could also be used for testing components theories. And it avoids the element of circularity in judging the roles of brain regions using the assumption that the brain has component parts specialized for particular purposes. (Since mapping of connected regions plus data about the effects of localized damage would show correlations between these extended regions and behaviour.) Einstein's suggestion, turned into a definite experiment by Perrin, that components of the molecular theory of gases and liquids could be tested by controlled Brownian motioncould be used to test the molecular theory independently of assumptions about either the existence of molecules and the distribution of momentum within them. It could also be used to isolate the bare molecular structure hypothesis from the need for correlations with thermodynamics, or the rather abstract but influential ideas that whatever is going on behind the gas laws at a microscopic level as an element of randomness.

accumulation

The limited scope for circularity allows experimental results and methods to accumulate, surviving changes of doctrine and improvements of theoretical technique. The experiments that were used to support both particle and wave theories of light can still be performed, even though we think that there are wrong assumptions behind the presuppositions and interpretations of both. They are often are performed as parts of school and university physics courses.

There are two very different but complementary sources of scientific accumulation. The first is the mass of past experiments, with their transitions from initiation to outcome, the skills developed in performing them, and the inferences needed to make sense of the data they yield. The other is the library of theory-constructing techniques, many of them mathematical, that can be transplanted from one topic to another. We can handle wave equations whether they are applied to electromagnetism or massive particles or traffic flow; we are used to extremal principles throughout dynamics; we know how to adapt computer models from climate prediction to population biology to economics. The techniques themselves are neutral. They are not true or false, supported or undermined by evidence, but are a resource for making things that are.[145]

The accumulation of experiments and the accumulation of theoretical techniques assist one another. Experiment provides relatively un-theorized material waiting for general formulation. When the library of techniques contains formulations, solutions, and

145 This is an extremely weak form of structural realism (highlighting the structural aspect but neutral about the realism aspect). It is (even) weaker than what Ladyman (2016) calls epistemic structural realism.

approximations that we have already seen how to transfer from one area to another, we know how they can represent the transitions from trigger to target in ways that model ranges of experiment without commitment to claims about what lies beneath these transitions. (It is not that we do not want to understand such underlying processes, but the more that experiments can be described as indicating rather than presupposing them the better.) The accumulation of theoretical devices thus facilitates the accumulation of autonomous experiments.[146]

sharing and attributing: the wide net

The characteristics of experiment that lead to accumulation also help epistemic cooperation involving a number of people with different skills. When individuals acting individually try to coordinate their inquiries they run up against differing interpretations of their experience, different understandings of the terms and their theories, and idiosyncratic degrees of belief. (What one person takes as established fact another may take as an interesting though dubious conjecture.) The accumulation of experimental results and their susceptibility to accumulating theoretical devices, though, gives a much more neutral place for shared activity. (They are much less conceptually sensitive.) The accumulating facts about an experiment are just how it was set up, physically, and what events it produced, and the theme of a free-ranging technique is that one can follow this kind of recipe to get this kind of result. Both of these can be easily communicated across differences of expectation, attitude, and theoretical vocabulary.

146 This is near to what Hacking (1992) calls ”stability”.

The resulting epistemic cooperation puts many brains to work on the same problems and combines many sources of data. Other capacities can be applied to shared projects for the same reasons. There is more that we can know this way. The result is knowledge because it is sensitive to the facts, and indeed there are more resources for sensitivity than when restricted to individual belief, since the responsiveness can be distributed among many co-workers.

the optimistic aspect

Tuneability, causation, separability non-circularity, accumulation, shared resources, suggestiveness: all major epistemic desiderata. These make experimentation our most powerful tool. (Just as they make the informal precursors of experimentation the most powerful source of non-scientific knowledge.)

There is a downside, though. The finely tuned experiments needed for this bonanza are not always possible, sometimes because we do not have enough control of the phenomena and sometimes because of limits in our capacities to design in advance and interpret afterwards. To run causally fine-grained experiments we need to produce fine-grained causal processes. That poses technological and theoretical problems. As a result, experimentation has to be part of a larger epistemic and technological context. A vital part, essential to the success of the whole project.

the sceptical part

Early epistemologies thought about scientific method in the context of general human strategies for acquiring knowledge and their prospects of success. They often contrasted the understanding of nature proffered by emerging science with the traditional lore of their cultures, which they deemed hard to justify and very often false. In the subsequent development of the subject, particularly after the middle of the twentieth century, this ambition was lost, although much was gained in terms of clarity and psychological realism. The task was split and passed to other disciplines. On the one hand philosophers of science thought about the characteristic methods of scientific inquiry. And on the other hand statisticians honed detailed advice about the structure of inquiries and the interpretation of data. Now no working scientist faced with a problem about how to investigate a topic will take her questions to an epistemologist. Her aim is to acquire knowledge but the theory of knowledge is largely irrelevant. I suspect that something similar is true of many issues in nonscientific inquiry.[147]

The role of statistics is not going to diminish. If anything it is going to increase, although the interpretation may change. But there is still a lot to say about how we think of evidence and its relation to the aims of inquiry, in particular to concepts such as knowledge which register the tightness of the connection between thought and fact. My line has been that the particular form our concept of knowledge takes is a special case of more general conditions of fit between what we think and the way things are. And in terms of these conditions there is room for an updated form of scepticism.[148] Much of what we believe, including central scientific doctrine, is not connected to the facts as well

147 Coady (2012).

148 It is a form of scepticism about knowledge rather than early modern scepticism which doubts that we have good reasons for our beliefs and then sometimes describes its conclusions in terms of "knowledge". Sorting out the two scepticisms would have been beyond our philosophical sophistication before the twentieth century

as we might hope. If we want improve the situation the most promising means is to expand our repertoire of experimental techniques. It is likely that more powerful experimental techniques will require experiment in the sociology of science also, as the number of people and the variety of distinct skills they draw on will surely become more and more intricate.[149] Skills honed by thinking about human knowledge are a real but very small part of the spectrum.

But there are no guarantees and in some cases the prospects are daunting. It is far from obvious that we will ever see how to make experiments to back up the explanatory power of the quark model of neutrons and protons. There is room for doubt that we will move beyond the standard model of elementary particles. Perhaps theories of gravitation and of the other fundamental forces will never be united. Perhaps the best we will be able to do in order to get a detailed grasp of how the human nervous system accomplishes complex tasks will be to simulate it with artificial systems whose detailed workings are equally inscrutable to us. Perhaps fundamental physics or brain science will require teams of experimenters that are too large and interact in ways that are too complex for our inflexible social capacities. The list of topics and reasons where progress might stall is long. And, the main point now, this scepticism about future progress is reinforced by an emphasis on experiment. Even if we can make satisfying theories of these things, with enough explanatory power and few enough plausible alternatives that we take ourselves to know them, we may be frustrated searching for the experiments

149 Knowledge has always been shared, and there has always been something deeply counterintuitive about purely individual knowledge. When people express sceptical rhetoric in everyday life they usually agitate about the possibility that "we" are subject to mass delusion, although on standard philosophical sceptical scenarios the existence of other people is as doubtful as that of the usual objects of thought. But for the person in the street deep error about others is even harder to grasp than deep error about the rest of the environment.

that give a firmer and more fruitful grasp of them. We may know without knowing as well as we want to.

The scepticism touches how we think about human knowledge as well as our capacities for achieving it. Reflection on experiment suggests that we can often achieve higher standards than those we are usually content called knowledge. But we often fall short of them, also. If we held ourselves to the standard of knowing as well as we do at our most successful then we would have to conclude that we often fail.

BIBLIOGRAPHY

AAAS. 2019. https://www.aaas.org/frances-arnolds-directed-evolution

Abbott, B. P.; et al. (LIGO Scientific Collaboration and Virgo Collaboration). 2016. "Observation of Gravitational Waves from a Binary Black Hole Merger," *Physical Review Letters* 116, 061102.

Achinstein, P. 2001. *The Book of Evidence*. New York: Oxford University Press.

Achinstein, P. (ed.). 2005. *Scientific Evidence: Philosophical Theories and Applications*. Baltimore, MD: John's Hopkins University Press.

Andersen, H. 2013. When to Expect Violations of Causal Faithfulness, and Why it Matters. *Philosophy of Science* 80, 672–683.

Anscombe, G. E. M. 1971. Causality and Determination: An Inaugural Lecture, Cambridge: Cambridge University Press. Reprinted in *The Collected Philosophical Papers of G. E. M. Anscombe*, Volume 2, Minneapolis, MN: University of Minnesota Press, 1981, 133–47.

Bailey, N. (1964). Elements of Stochastic Processes. New York: Wiley.

Bailey, R. A. (2008) *Design of Comparative Experiments*. Cambridge UK: Cambridge University Press.

Ballargeon, R., Spelke, E. S., and Wasserman, S. 1985. "Object Permanence in Five Month-Old Infants," *Cognition* 20, 191-208.

Balcolmbe, J. 2016. *What a Fish Knows*. Farrar, Straus and Giroux.

Bandyopadhyay, P. and Foster, M. R., (eds.) 2013. *Handbook of the Philosophy of Statistics*. Amsterdam: Elsevier.

Bandyopadhyay, P. S., Brittan Jr., Gordon, and Taper, M. L. 2011. *Belief, Evidence, and Uncertainty: Problems of Epistemic Inference.* New York: Springer.

Bennett, D. (2013) "Defining Randomness," in Bandyopadhyay and Forster (2013), 683-639.

Bennett, J. 2003. *A Philosophical Guide to Conditionals*. Oxford: Clarendon Press.

Bernardo, J. M. 2013. “Modern Bayesian Inference: Foundations and Objective Methods,” in Bandyopadhyay and Forster (2013), 263-306.

Biener, Z. and Schliesser, E. (eds.). 2014. *Newton and Empiricism*. New York: Oxford University Press.

Birnbaum, A. 1972. “More on Concepts of Statistical Evidence,” *Journal of the American Statistical Association*. 67, 858-861.

Birnbaum, A. 1969. “Concepts of Statistical Evidence,” in Sidney Morgenbesser, Patrick Suppes, and Morton White (eds.), *Philosophy, Science and Method: Essays in Honor of Ernest Nagel*, New York: St. Martin's Press.

Bland, J. M., and D. G. Altman. 1996. “Statistics Notes: Measurement Error,” *British Medical Journal* 313, 744.

Borenstein, A. R., C. I. Copenhaver, and J. A. Mortimer. 2006. “Early-Life Risk Factors for Alzheimer Disease,” *Alzheimer Disease and Associated Disorders* 20, 63-72.

Braham, M. and M. van Hees. 2009. “Degrees of Causation,” *Erkenntnis* 71, 323–344.

Braithwaite, R. B. 1953. *Scientific explanation: A Study of the Function of Theory, Probability and Law in Science*. Cambridge: Cambridge University Press.

Brown, H. I. 1987. “Naturalizing Observation,” in N. Nersessian (ed.), *The Process of Science.* Dordrecht: Martinus Nijhoff Publishers, 179–194.

Brown, J. 2004. *Anti-Individualism and Knowledge*. Cambridge, MA: MIT Press.

Buckley, M. 2016. The Search for New Physics at CERN http://bostonreview.net/books-ideas/matthew-buckley-physics-series

Bulmer, M. G. 1979. *Principles of Statistics.* New York: Dover

Burge, T. 1988. “Individualism and Self-knowledge,” *Journal of Philosophy* 85, 649-63.

Camerer, C. F. 2003. *Behavioral Game Theory.* New Jersey: Princeton University Press.

Cartwright, N. 1983. *How the Laws of Physics Lie.* Oxford: Oxford University Press.

Cartwright, N. 1999. “Models and the Limits of Theory: Quantum Hamiltonians and the BSC Models of Superconductivity,” in M. Morrison, and M. S. Morgan, (eds.), *Models as Mediators: Perspectives on Natural and Social Science.* New York: Cambridge University Press.

Cartwright, N. 2004. “Causation: One Word, Many Things,” *Philosophy of Science*, 71, 805-819.

Cartwright, Nancy. 2010. “What Are Randomised Controlled Trials Good For?” *Philosophical Studies*, 147, 59-70.

Chabris, C. and D. Simons. 2010. *The Invisible Gorilla: and Other Ways Our Intuitions Deceive Us*. New York: Crown Publishers.

Christensen, D. 1983. "Glymour on Evidential Relevance," *Philosophy of Science* 50, 471-481.

Christensen, D. 1997. "What is Relative Confirmation?" *Nous* 31, 370-384.

Christensen, David. 1994. "Conservatism in Epistemology," *Noûs*, 28, 69-89.

Christensen, David. 1999. "Measuring Confirmation," *Journal of Philosophy* 96, 437-461.

Christensen, David 2010. "Higher-Order Evidence," *Philosophy and Phenomenological Research* 81, 185–215.

Coady, D. 2012. *What to Believe Now: Applying Epistemology to Contemporary Issues*. Chichester, UK: Wiley-Blackwell.

Cobb, M. 2015. *Life's Greatest Secret: The Race to Crack the Genetic Code*. New York: Basic Books.

Tal, E. and J. Comesaña. 2017. "Is Evidence of Evidence Evidence?" *Noûs* 51**,** 195-112.

Earman, J. 1992. *Bayes or Bust: A Critical Examination of Bayesian Confirmation Theory.* Cambridge, MA: MIT Press.

Cohen, J. 1994. "The earth is round (p<.05)," *American Psychologist*, 49, 997-1003.

Cook, T. and D. T. Campbell. 1979. *Quasi-experimentation: Design and Analysis for Field Settings*. Boston: Houghton Mifflin.

Cox, D. 1986. "Statistics and Causal Inference: comment on Holland," *Journal of the American Statistical Association* 81. 963–964.

Cox, D. 1990. "The Role of Models in Statistical Analysis," *Statistical Science* 5, 169-174.

Craig, William. 1990. *Knowledge and the State of Nature*. Oxford: Oxford University Press.

Davey Smith, G. and S. Ebrahim .2003. "'Mendelian Randomization': Can Genetic Epidemiology Contribute to Understanding Environmental Determinants of Disease?" *International Journal of Epidemiology* 32, 1–22.

Dawid, A.P. 2000. "Causal Inference without Counterfactuals," *Journal of the American Statistical Association* 95, 407-424.

Dawid, A. P., M. Musio, and S. E. Fienberg. 2019. "From Statistical Evidence to Evidence of Causality," *Bayesian Analysis* 11, 725–752.

Dingfelder, S. F. 2010. "The First Modern Psychology Study," *American Psychological Association Monito*r 41, 30.

Dowe, P. 2000. *Physical Causation*. Cambridge: Cambridge University Press.

Dretske, F. 1973. "Contrastive Statements," *Philosophical Review* 81, 411-437.

Duhem, P. (1905) *La Théorie Physique: Son Objet, Sa Structure*. Paris: Marcel Rivière; 2nd edition, 1914.

Earman, J. 1992. Bayes or Bust: A Critical Examination of Bayesian Confirmation Theory. Cambridge, MA: MIT Press.

Eberhardt, F. and R. Scheines. (2007). "Interventions and Causal Inference," *Philosophy of Science*, 74, 981–995.

Ekstrom, P., and D. Wineland. (1980). "The Isolated Electron," *Scientific American* 243, 104-121.

Elgin, C. Z. 2004. "True Enough," *Philosophical Issues* 14, 113–131.

Elgin, C. Z. 2008. "Trustworthiness," *Philosophical Papers* 27: 371-387.

Faulkner, P. 2000. "Testimonial Knowledge," *Journal of Philosophy* 97, 581-601.

Field, H. 2009. "Epistemology Without Metaphysics," *Philosophical Studies* 143: 249-90.

Fisher, R. A. 1935. *The Design of Experiments.* Edinburgh and London: Oliver and Boyd.

Forster, M., and E. Sober. 1994. "How to Tell when Simpler, More Unified, or Less Ad Hoc Theories will Provide More Accurate Predictions," *British Journal for the Philosophy of Science.* 45, 1-35.

Fraser, L. H. and P. Keddy. 1997. "The Role of Experimental Microcosms in Ecological Research," *Trends in Ecology and Evolution* 12, 478–481.

Franklin, A. 1994. "How to Avoid the Experimenters' Regress," *Studies in History and Philosophy of Science* 25, 463-491.

Franklin, A. 1986. *The Neglect of Experiment*. New York: Cambridge University Press.

Franklin, A. 2016. *What Makes a Good Experiment? Reasons and Roles in Science.* Pittsburgh, PA: University of Pittsburgh Press.

Franklin , A., A.W.F Edwards, D. Fairbanks, and D. Hartl. 2008. *Ending the Mendel-Fisher Controversy*. Pittsburgh, PA: University of Pittsburg Press.

Franklin, A. and Perovic, S. 2016. "Experiment in Physics," *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2016/entries/physics-experiment/>.

Freedman, D. and P. Humphreys. 1999. "Are There Algorithms That Discover Causal Structure?" *Synthese* 121, 29-54.

Galison, P. 1987. *How Experiments End*. Chicago: University of Chicago Press

Galison, P. 1997. *Image and Logic: A Material Culture of Microphysics*. Chicago: University of Chicago Press.

Gandenberger, G. 2014. "A New Proof of the Likelihood Principle," *British Journal for the Philosophy of Science 66.* 475-503.

Garthwaite, P., I. Jolliffe, and B. Jones. 2002. *Statistical Inference.* New York: Oxford University Press.

Gigerenzer, G. 2004. "Mindless Statistics," *The Journal of Socio-Economics* 33: 587–606.

Ginet, C. 1992. "Causal Theories in Epistemology," in J. Dancy and E. Sosa, (eds.), *A Companion to Epistemology.* Oxford: Wiley-Blackwell.

Glymour, C. 1980. *Theory and Evidence*. New Jersey: Princeton University Press.

Glymour, C. 1986. "Statistics and Causal inference: comment on Holland," *Journal of the American Statistical Association* 81, 965–966

Glymour, C. 2001. *The Mind's Arrows: Bayes Nets and Graphical Causal Models in Psychology*. Cambridge, Mass: MIT Press.

Glymour, C., and G. F. Cooper (eds.). 1999. *Computation, Causation, and Discovery.* Cambridge, Mass: MIT Press.

Goldman, Alvin. 1988. *Epistemology and Cognition.* Second edition. Cambridge MA: Harvard University Press.

Gooding, D., T. Pinch, and S. Shaffer. 1989. *The Uses of Experiment*: *Studies in the Natural Sciences*. New York: Cambridge University Press.

Gopnik, A., C. Glymour, D. M. Sobel, L. E. Schulz, and T. Kushnir. 2004. "A Theory of Causal Learning in Children: Causal Maps and Bayes Nets," *Psychological Review* 111, 3–32.

Granger, C. 1986. "Statistics and causal inference: comment on Holland," *Journal of the American Statistical Association* 81. 967–968.

Greenland, S. 2011. "The Logic and Philosophy of Causal Inference: A Statistical Perspective," In Bandyopadhyay and Forster (eds.). 2011, 813-830.

Grossman, J. 2013. "The Likelihood Principle," in Bandyopadhyay and Forster (eds.). 2013, 553-580

Grossman J., and F. Mackenzie. 2005. "The Randomised Controlled Trial: Gold Standard, or Merely Standard?" *Perspectives in Biology and Medicine* 48, 516—534.

Guala, F. 2005. *The Methodology of Experimental Economics.* New York: Cambridge University Press.

H.O. [identified only by initials]. 1962. Scott and Scurvy. *Journal of the Canadian Medical Association* 87, 32-33

Hacking, I. 1965. *The Logic of Statistical Inference.* Cambridge UK: Cambridge University Press.

Hacking, I. 1983. *Representing and Intervening*. Cambridge UK: Cambridge University Press.

Hacking, I. 1992. "The Self-Vindication of the Laboratory Sciences," in A. Pickering (ed.), *Science as Practice and Culture.* Chicago: University of Chicago Press, 1992, 29-64.

Hájek, A. 2009. "Fifteen Arguments against Hypothetical Frequentism," *Erkenntnis* 70, 211-235.

Harman, G. 1965. "The Inference to the Best Explanation," *Philosophical Review* 74, 88-95.

Harman, G. 1968. "Knowledge, Inference, and Explanation," *American Philosophical Quarterly* 5, 88-95.

Harman, G. 1986. Change in View. Cambridge, MA: MIT Press.

Harper, W. 2012. *Isaac Newton's Scientific Method:Turning Data into Evidence about Gravity and Cosmology*. New York: Oxford University Press.

Hawthorne, J. 2005. "Degree-of-Belief and Degree-of-Support: Why Bayesians Need both Notions," *Mind* 114, 278-287.

Hetherington, S. 2002. *Good Knowledge, Bad Knowledge*: *On Two Dogmas of Epistemology.* New York: Oxford University Press.

Hetherington, S. (ed.). 2006. *Epistemology Futures*. New York: Oxford University Press.

Hawkins, D. 2014. *Biomeasurement: A Student's Guide to Biological Statistics*. Oxford: Oxford University Press.

Hawthorne, J. 2005. "Degree of Belief and Degree of Support: Why Bayesians Need Both," *Mind* 114, 277–320.

Hitchcock, C. 1995. "The Mishap at Reichenbach Fall: Singular vs. General Causation," *Philosophical Studies* 78, 257-291.

Hitchcock, C. 1996. "The Role of Contrast in Causal and Explanatory Claims," *Synthese* 107, 395–419.

Hookway, C. 2003. "How to Be a Virtue Epistemologist," in M. DePaul and L. Zagzebski (eds.), Intellectual Virtue: Perspectives from Ethics and Epistemology. New York: Oxford University Press, 183-202.

Halpern, J. Y. and J. Pearl. 2005. "Causes and Explanations: A Structural-Model Approach. Part I," *The British Journal for the Philosophy of Science* 56, 843-887.

Howson, C., and Urbach, P. 1989. *Scientific Reasoning: the Bayesian Approach.* La Salle, Ill.: Open Court.

Hock, R. R. 2002. *Forty Studies that Changed Psychology*. Upper Saddle River: Prentice-Hall.

Holland, P. 1986. "Statistics and Causal Inference," *Journal of the American Statistical Association* 81, 945–70. Rejoinder, 968–970.

Holman, B. 2015. "Why Most Sugar Pills Are Not Placebos," *Philosophy of Science* 82, 1330-1343.

Howell, D. (web). https://www.uvm.edu/~dhowell/StatPages/ResamplingWithR/ResamplingR.html

Hsu, Ping-I, Pei-Chin Lin, D. Y Graham. 2015. "Hybrid Therapy for *Helicobacter pylori* Infection: A Systemic Review and Meta-analysis," *World Journal of Gastroentero*logy 21, 12954-12962.

Humphreys, P. 2014. *The Chances of Explanation: Causal Explanation in the Social, Medical, and Physical Sciences.* New Jersey: Princeton University Press.

Humphreys K., J. C. Blodgett, and T. H. Wagner. 2014. "Estimating the Efficacy of Alcoholics Anonymous without Self-Selection Bias: An Instrumental Variables Re-Analysis of Randomized Clinical Trials," *Alcoholism: Clinical Experimental Research* 38, 2688–2694.

Humphreys, W. C. 1967. "Galileo, Falling Bodies and Inclined Planes: An Attempt at Reconstructing Galileo's Discovery of the Law of Squares," *The British Journal for the History of Science* 3, 225-244

Hunt, M. 1997. *How Science takes Stock: the Story of Meta-Analysis*. New York: Russell Sage Foundation.

Hunter, J., and F. Schmidt. 2004. *Methods of Meta-analysis: Correcting Error and Bias in Research*. Thousand Oaks, CA: Sage Publications

Jenkins Ichikawa, J. 2017. *Contextualising Knowledge: Epistemology and Semantics*. New York: Oxford University Press.

Illari, P. and F. Russo. 2014. *Causality: Philosophical Theory Meets Scientific Practice.* New York: Oxford University Press.

Imbens, G. and D. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. New York: Cambridge University Press.

Jaggard, V. 2016. "Five Small Animals that Messed with Big Science," http://news.nationalgeographic.com/2016/04/160429-lhc-weasel-animals-science-machines.html

James, W. 1897. *The Will to Believe and Other Essays in Popular Philosophy*. London: Longmans, Green & Co.

Johnson, D. 1991. "Induction and Modality," *Philosophical Review* 50, 399-430.

Karjalainen, A. and A. Morton. 2003. "Contrastive knowledge," *Philosophical Explorations,* 6, 74-89.

Kahneman, D. and A. Tversky. 1979. "Prospect Theory: An Analysis of Decision Under Uncertainty," *Econometrica* 47, 263-291.

Kawecki, T., et al. 2012. "Experimental Evolution," *Trends in Ecology and Evolution*, 27, 547-560.

Kerr, E., and A. Gelfert. 2014. "The 'Extendedness' of Scientific Evidence," *Philosophical Issues* 24, 253-281.

Kimball, A. W. 1957. "Errors of the Third Kind in Statistical Consulting," *Journal of the American Statistical Association* 52, 133–42.

Kirsh, D. and P. Maglio. 1994. "On Distinguishing Epistemic from Pragmatic Action," *Cognitive Science* 18, 513–49.

Kitcher, P. 1990. "The Division of Cognitive Labor," *Journal of Philosophy* 87, 5-22.

Kitcher, P. 1995. *The Advancement of Science*. New York: Oxford University Press.

Kramer, P. D. 2016. *Ordinarily Well: The Case for Antidepressants*. New York: Farrar, Straus and Giroux.

Kravitz, D. J. and M. Behrmann. 2011. "Space-, Object-, and Feature-Based Attention Interact to Organize Visual Scenes," *Attention, Perception & Psychophysics, 73*, 2434-2447.

Kreps, D. M. 1990. *Game Theory and Economic Modelling.* New York: Oxford University Press.

Kripke, S. 1980. *Naming and Necessity.* Cambridge, MA: Harvard University Press.

Kripke, S. 2011. "Two Paradoxes of Knowledge," in S. Kripke, *Philosophical Troubles: Collected Papers (Volume 1*). New York: Oxford University Press, 27–51.

Lackey, J. 2014. "Socially Extended Knowledge," *Philosophical Issues* 24, 282-298.

Ladyman, J. 2016. "Structural Realism," *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2016/entries/structural-realism/>.

Lehmann, E.L. and J. Romano. 2005. *Testing Statistical Hypotheses*. 3rd Edition. New York: Springer Science+Business Media, LLC.

Leigh, J. P. and M. Schembri. 2004. "Instrumental Variables Technique: Cigarette Price Provided better Estimate of Effects of Smoking on SF-12," *Journal of Clinical Epidemiology* 57, 284–293

LeLorier, J., G. Gregoire, A. Benhaddad, J. Lapierre, and F. Derderian. 1997. "Discrepancies Between Meta-Analyses and Subsequent large Randomized, Controlled Trials," *New England Journal of Medicine* 337, 536–543.

Levi, Isaac. 1974. *Gambling with Truth.* Cambridge, MA: MIT Press

Lewis, D. 1986. "Causation," in *Philosophical Papers*, Vol. II. New York: Oxford University Press, 159-171.

Lewis, D. 1973. *Counterfactuals*. Oxford: Basil Blackwell.

Lewis, D. 1996. "Elusive Knowledge," *Australasian Journal of Philosophy* 74, 549-67.

Lewis, David. 2000. "Causation as Influence," *Journal of Philosophy* 97, 182-197.

Li, D. 2011. "Diabetes and Pancreatic Cancer," *Molecular Carcinogenesis* 51, 64-74.

Lipton, P. 1991. *The Inference to the Best Explanation*. London: Routledge.

Longair, M. S. 2006. *The Cosmic Century: A History of Astrophysics and Cosmology.* Cambridge UK: Cambridge University Press.

Lycan, W. G. 2005. "Explanation and Epistemology," in P. K. Moser (ed.), *The Oxford Handbook of Epistemology,* New York: Oxford University Press, 408-433.

Mackie, J.L. 1965. Causes and Conditions. *American Philosophical Quarterly* 2, 245-264.

Marshall B.J. and J. R. Warren. 1984. Unidentified Curved Bacilli in the Stomach of Patients with Gastritis and Peptic Ulceration," *The Lancet* 323, 1311–5.

Mayo, D. 1996. *Error and the Growth of Experimental Knowledge*. Chicago: Chicago University Press.

Mayo, D. 2005. Evidence as Passing Severe Tests. In Achinstein, 2005, 95-127.

Mayo, D. 2018*. Statistical Inference As Severe Testing – How to Get Beyond the Statistics Wars.* New York: Cambridge: Cambridge University press

Mayo, D. & A. Spanos. 2013. "Error Statistics," in Bandyopadhyay and Forster (2013), 153-198.

McCullagh, P. 2002. What is a Statistical Model? *Annals of Statistics* 30. 1225–131

Meek, C. and C. Glymour. 1994. Conditioning and Intervening. *British Journal for the Philosophy of Science*, 45, 1001-1021

Mellor, H. 1971. *The Matter of Chance.* Cambridge UK: Cambridge University Press.

Memetea, S. 2015. *Simpson's Paradox Unified*. PhD Thesis, UBC.

Meselson, M. & F. W. Stahl. 1958. "The Replication of DNA in *Escherichia Coli*," PNAS 44, 671–82.

Michaelian, K. 2014. "From Distributed Concepts to Distributed Reliabilism," *Philosophical Issues* 24, 314–346.

Mill, J. S. 1843. *A System of Logic.* Eight Edition. New York: Harper and Brothers.

Miyake, T. 2013a. "Underdetermination, Black Boxes, and Measurement," *Philosophy of Science* 80, 697-708

Miyake, T. 2013b. "Essay Review: Isaac Newton's Scientific Method," *Philosophy of Science* 80, 310-316.

Mogil, J.S., and M.L. Chanda. 2005. "The Case for the Inclusion of Female Subjects in Basic Science Studies of Pain, *Pain* 1-2, 1-5.

Mole, C. 2015. "Attention and Cognitive Penetration," in J. Zembeikis and T. Raftopoulos (eds.), *The Cognitive Penetrability of Perception*, Oxford: Oxford University Press, 218-238.

Morgan, M. S. and M. Morrison. (eds.). (1999) *Models as Mediators*. New York: Cambridge University Press.

Morton, A (1986)review of Ian Hacking Representing and Intervening, *Phil. Rev*.,**95**,4: 606-611.

Morton, A. 2002. *The Importance of Being Understood: Folk Psychology as Ethics*. London: Routledge.

Morton, A. 2010. "Contrastivism," in D. Pritchard and S. Bernecker (eds.), *The Routledge Companion to Epistemology.* London: Routledge, 513-522.

Morton, A. 2012a. *Bounded Thinking: Intellectual Virtues for Limited Agents* Oxford University Press.

Morton, A. 2012b. "Contrastive Knowledge", in M. Blauuw (ed.) *Contrastivity in Philosophy.* London: Routledge

Morton, A. 2012c. "Accomplishing Accomplishment," *Acta Analytica.* 27, 1-8

Morton, A. (2013). Accomplishment. Direct posting to *Phil Papers* http://philpapers.org/rec/MORA-9

Morton, A. 2014a. "Acting to Know," in A. Fairweather (ed.), *Virtue Epistemology Naturalized: Bridges between Virtue Epistemology and Philosophy of Science.* Springer: Synthese Library 366, 195-207.

Morton, A. 2014b. Shared Knowledge from Individual Vice: The Role of Unworthy Epistemic Emotions. *Philosophical Inquiries* 2, 163-172.

Morton, A. 2016. Review of Allan Franklin (2016), Metascience 102

Morton, A. and N. Cockram. 2017. "Contrastivism," in *Oxford Bibliographies Online.*

Morton, A. and A. Karjalainen. 2003. "Contrastive knowledge," *Philosophical Explorations,* 6. 74-89.

Morton, A. and A. Karjalainen. 2008. "Contrastivity and Indistinguishability," *Social Epistemology* 22, 271–280.

Mößner, N. and A. Nordmann. (eds.). (2017). *Reasoning in Measurement*. London: Routledge.

Myrvold, W. and W. Harper. 2002. "Model Selection, Simplicity, and Scientific Inference," *Philosophy of Science* 69, S135-S149.

Splawa-Neyman, J., D. M. Dabrowska, and T. P. Speed. (1923/1990). "On the Application of Probability Theory to Agricultural Experiments: Essay on Principles," *Statistical Science* 5, 465-472.

Neyman, J. 1957. "Inductive Behavior as a Basic Concept of Philosophy of Science," *Revue de l'Institut International de Statistique / Review of the International Statistical Institute* 25, 7-22

Nisbett, Richard. 2016. The Crusade Against Multiple Regression Analysis. Edge.Org.

Nguyen, L., Hua, H. and Chuong P.-H. (2006). "Chiral Drugs: An Overview," *International Journal of biomedical science* 2, 85-100.

Nozick, R. 1981. *Philosophical Investigation*s. Cambridge MA: Harvard University Press.

O. H. see H.O.

Ohsumi, Y., et al. 1992. "Autophagy in Yeast Demonstrated with Proteinase-Deficient Mutants and Conditions for Its Induction," *The Journal of Cell Biology* 119, 301–311.

Pawitan, Y. 2013. *In All Likelihood: S statistical Modelling and Inference Using Likelihood*.

New York: Oxford University Press.

Pearl, J. 2000. *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press.

Pearl, J., M. Glymour, and N. Jewell. 2016. *Causal Inference in Statistics: A Primer*. New York: Wiley

Perini, L. 2017. "The Altered Image: Composite Figures and Evidential Reasoning with Mechanically Produced Images," in Mößner and Nordman (2017), 87–98.

Perovic, S. 2011. "Missing Experimental Challenges to the Standard Model of Particle Physics," *Studies in History and Philosophy of Modern Physics* 42, 32–42.

Perovic, S. 2017. "Experimenter's Regress Argument, Empiricism, and the Calibration of the Large Hadron Collider," *Synthese* 194, 313–332.

Poirotte, C., et al. 2016. "Morbid Attraction to Leopard Urine in Toxoplasma Infected Chimpanzees," *Current Biology* 26, R83–R101.

Pritchard, D. 2009. Safety-Based Epistemology: Whither Now? *Journal of Philosophical Research* 34. 33-45.

Putnam, Hilary. 1973. Meaning and Reference. *The Journal of Philosophy* 70, 699-711.

Quine, W.V. 1953. Two Dogmas of Empiricism. In *From a Logical Point of View*. Cambridge, MA: Harvard University Press, 20-46.

Rabinowitz, D. web. The Safety Condition for Knowledge. *Internet Encyclopedia of Philosophy* http://www.iep.utm.edu/safety-c/

Radder, H. 1988. *The Material Realization of Science*. Assen: Van Gorcum.-online

Radder, H. (ed.) 2003. *The Philosophy of Scientific Experimentation*. Pittsburgh, PA: University of Pittsburgh Press

Randall, Lisa (2011) *Knocking on Heaven's Door. Ecco.*

Rheinberger, Hans-Jörg (1997) *Toward a history of Epistemic Things : Synthesizing Proteins in the Test Tube.* Stanford: Stanford University Press.

Rubin, Donald B. (1974) Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology* 66, 688-701

Rubin, D (1986) Statistics and causal inference: Which if's have causal answers, comment on Holland. *Journal of the American Statistical Association* 81. 961–962

Rutherford, E. 1911. "The Scattering of α and β Particles by Matter and the Structure of the Atom". *Philosophical Magazine*. Series 6. 21, 669–688.

Ruxton, G. and N. Colegrave. 2003. *Experimental Design for the Life Sciences*. New York: Oxford University Press.

Saey, T. 2015. "Scientists Tackle the Irreproducibility Problem," *Science News* 188, 22

Schaffer, J. 2005. Contrastive Knowledge. In Tamar Szabo Gendler & John Hawthorne (eds.), *Oxford Studies in Epistemology 1*. Oxford: Oxford University Press, 235

Schaffer, J. (2000) Causation by Disconnection. *Philosophy of Science*, 67, 285–300.

Schaffer, J. (2005) Contrastive Causation. *Philosophical Review*, 114, 327-358. .

Schluter, D. 2000. *The Ecology of Adaptive Radiation*. New York: Oxford University Press.

Smith, G. E. 2002. "The Methodology of the Principia," in *The Cambridge Companion to Newton*, I. B. Cohen and G. E. Smith (eds.), Cambridge: Cambridge University Press, 138-173

Smith, G. E. 2014. "Closing the Gap: Testing Newtonian Gravity, Then and Now," In *Newton and Empiricism*, Z. Biener and E. Schliesser (eds). New York: Oxford University Press, 262-351.

Sober, E. 2001. "Venetian Sea Levels, British Bread Prices, and the Principle of the Common Cause," *The British Journal for the Philosophy of Science*, 52, 331-346.

Pritchard, D. 2015. "Anti-Luck Epistemology and the Gettier Problem," *Philosophical Studies* 17, 93-111.

Quine, W. V. 1981. "Success and Limits of Mathematization," in *Theories and Things,* Cambridge, MA: Harvard University Press: 148–155.

Rabinowitz, D. (web). "The Safety Condition for Knowledge," *Internet Encyclopedia of Philosophy* (https://www.iep.utm.edu/safety-c/)

Sosa, E. 1999. How Must Knowledge be Related to What is Known? *Philosophical Topics* 26, 373-384.

Spirtes, P., C. Glymour, R. Scheines. 2001. *Causation, Prediction, and Search*. Cambridge MA: M.I.T. Press.

Staley, K. 1996. "Novelty, Severity, and History in the Testing of Hypotheses: the Case of the Top Quark," *Philosophy of Science* 6, S248-S255

Staley, K. 2004. "Robust Evidence and Secure Evidence Claims," *Philosophy of Science* 71, 467–488.

Staley, K. 2004. *The Evidence for the Top Quark: Objectivity and Bias in Collaborative Experimentation.* New York: Cambridge University Press.

Stalnaker, R. C. (1968). "A Theory of Conditionals," in N. Rescher (ed.), *Studies in Logical Theory (American Philosophical Quarterly Monographs 2)*. Oxford: Blackwell, 98-112.

Stegenga, J. 2018. *Medical Nihilism*. Oxford: Oxford University Press.

Steyn, D. G., and S. Galmarini. 2008. "Evaluating the Predictive and Explanatory Value of Atmospheric Numerical Models: Between Relativism and Objectivism," *The Open Atmospheric Science Journa*l 2008, 2, 38-45.

Strevens, M. 2011. "Probability out of Determinism," In C. Beisbart and S. Hartmann (eds.), *Probabilities in Physics*. New York: Oxford University Press, 339--364

Strevens, M. (2013) *Tychomancy: Inferring Probability from Causal Structure. Cambridge, MA.:* Harvard University Press,

Stroebel, W., and F. Strack. 2014. "The Alleged Crisis and the Illusion of Exact Replication," *Perspectives on Psychological Science* 9, 59–71

Sussman J. B., Hayward R. 2010. "An IV for the RCT: Using Instrumental Variables to Adjust for Treatment Contamination in Randomized Controlled Trials," *British Medical Journal* 340, 1181–1184.

Tal, E. and J. Comesaña. 2017. "Is Evidence of Evidence Evidence?" *Noûs* 51**,** 195-112.

Toft, C. A. And T. F. Wright. (2015) *Parrots of the Wild.* Berkeley: University of California Press

Van Bavel, J. J., J. J. Van Bavel, P. Mende-Siedlecki, W. J. Brady, and D. A. Reinero. 2016. Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences* 113, 6454-6459

van Fraassen, B. C. 1983. "Theory Comparison and Relevant Evidence," in J. Earman (ed.), *Testing Scientific Theories. Minnesota Studies in the Philosophy of Science, Volume 10.* Minneapolis: University of Minnesota Press, 27-42.

Warner, R. R. 1988. "Traditionality of Mating-Site Preferences in a Coral Reef Fish," *Nature* 335, 719-721.

Wasserman, L. 2004. *All of Statistics: a Concise Course in Statistical Inference.* New York: Springer.

Wasserstein, R. L. and N. A. Lazara. 2016. "The ASA's statement on *p*-Values: Context, Process, and Purpose," *The American Statistician* 70, 129-133.

Weber, E. 2009. "How Probabilistic Causation Can Account for the Use of Mechanistic Evidence," *International Studies in the Philosophy of Science* 23, 277–295

Weber, M. 2005. *Philosophy of Experimental Biology. New York:* Cambridge University Press.

Welbourne, M. 1986. *The Community of Knowledge*. Aberdeen: Aberdeen University Press.

Wimmer, H. and J. Perner. 1983. "Beliefs about Beliefs: Representation and Constraining Function of Wrong Beliefs in Young Children's Understanding of Deception," *Cognition* 13, 103–128.

Will, C. 2018. "General Relativity Verified by a Triple-Star System," *Nature,* news and views. 4 July 2018, on web at /www.nature.com/articles/d41586-018-05549-4.

Williamson, T. 2000. *Knowledge and its Limits*. New York: Oxford University Press.

Wilson, M. 2017. *Physics Avoidance*. New York: Oxford University Press.

Woodward, J. 2011. "Data and Phenomena: A Restatement and a Defense," *Synthese* 182, 165–179

Woodward J. (2006). "Sensitive and Insensitive Causation", *Philosophical Review* 115, 1-50.

Woodward, J. (2003a) "Experimentation, Inference, and Causal Realism," in Radder (ed.), (2013), 87-118.

Woodward, J. 2003b. *Making Things Happen*. New York: Oxford University Press.

Woodward, J. 2016a. "Causation and Manipulability," *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2016/entries/causation-mani/>.

Woodward, J. 2016b. "The Problem of Variable Choice," *Synthese* 193, 1047–1072.

Woodward, J. and J. Bogen. 1988. "Saving the Phenomena," *Philosophical Review* 97, 303-352.

Wray, B. 2002. "The Epistemic Significance of Collaborative Research," *Philosophy of Science* 69, 150–168.

Wu, C. F. J., and M. Hamada. 2002. *Experiments: planning, analysis, and Parameter Design Optimization*. New York: Wiley.

Xu, V., M. Jaffe, C. D. Panda, S. L. Kristensen, L. W. Clark, Müller, H. 2019. "Probing Gravity by Holding Atoms for 20 Seconds," *Science* 366, 745-749.

Zagzebsky, L. 1996. *Virtues of the Mind*. New York: Cambridge University Press.

Ziegler, V. L. 2004. *Trial by Fire and Battle in Medieval German Literature*. Rochester: Camden House.

Zwier, K. (to appear). Interventionist Causation in Thermodynamics. *Philosophy of Science*.

Zwier, K. 2013. "An Epistemology of Causal Inference from Experiment," *Philosophy of Science* 80, 660-671.