

## **PART I: MATTER**

**In these two chapters I sketch the basic principles of a dialectical materialist understanding of nature, and consider its implications for some of the key issues in the philosophy of science, including emergence, supervenience, prediction, explanation, and induction. The main target of chapter 2 is reductionism, and the main target of chapter 3 is empiricism.**

## Chapter 2

### Anti-Reductionism

---

There is an old bit of advice which says: Watch your friends; your enemies will take care of themselves. In the scientific *métier*, this saying goes: Suspect the obvious; the obscure truths will elude you anyway.

— Kline, *Mathematics in Western Culture*

The sting is only removed from a system of thought when the particular conditions under which it makes sense are described.

— Bhaskar, *A Realist Theory of Science*

Over the last three hundred years the scientific materialist understanding of the world has been moulded by two pictures, two *intuitions*, and I want to convince you that a third is at least possible. The first intuition is *reductionism*. Societies are made of people, people are made of cells, cells of molecules, and molecules of atoms. These parts come together to form wholes, and the behaviour of the wholes are determined by the parts. The logical conclusion of reductionism is that all science — and, in extreme cases, all art, ethics, and politics — is just ‘physics plus abbreviations’ as the logical positivists put it.

Reductionism has been one the most powerful ideas of the modern age, but many are reluctant to reach its seemingly brutal conclusions. ‘Indeed, it seems to be a little-known law governing the behaviour of contemporary philosophers that whenever they profess faith in any form of materialism or physicalism they must make it absolutely clear that they are, of course, in no way endorsing anything as unsophisticated, reactionary, and generally intolerable as reductionism’ (Melnyk, 1995).

The alternative intuition is less of a militantly cohesive picture than reductionism, and more of an understandable reaction to it. Let us call it *pluralism*, though the same intuition goes under different names in different contexts. The central point is that objects — including people, and their art ethics, and politics — must be understood on their own terms. As Fodor, in a review of E.O.Wilson’s reductionist manifesto *Consilience: The Unity of Knowledge*, puts it

everything is physical perhaps, but surely there are many kinds of physical things. Some are protons; some are constellations; some are trees or cats; and some are butchers, bakers or candlesticks. For each kind of thing, there are the proprietary

generalisations by which it is subsumed, and in terms of which its behaviour is to be explained. For each such generalisation there is the proprietary vocabulary that is required in order for our discourse to express it. Nothing can happen except what the laws of physics permit, of course; but much goes on that the laws of physics do not talk about.

It is important to realise that there is no necessary contradiction between pluralism and reductionism, and most modern philosophies combine elements of both. For example, many scientists and philosophers are increasingly unhappy with reductionist claim that the higher levels are *determined* by the lower, and one attractive alternative is what we may call *pragmatic anti-reductionism*. This argues that although reductionism may be correct in principle, it can rarely be used in practice: it is simply not feasible to collect all the data, and perform the calculations necessary, for all but the most trivial systems. According to pragmatic anti-reductionism, properties of wholes may be determined by those of the parts, but this does not imply that they are necessarily derivable from them. If you want to understand the world then the only possible strategy is to investigate each phenomenon on its own terms rather than start from the physics. In my experience most practising scientists would agree with some form of pragmatic anti-reductionism.

Basic pragmatic anti-reductionism can be strengthened in various ways. For example, we can borrow from chaos theory and argue that aggregate properties of the system may be sensitive to some properties of a part, such as the infamous sensitivity of weather systems to a butterfly's wing. If this is the case then accurate predictions about the higher level depend on knowing the properties of the parts with unbounded accuracy, and there are various reasons, such as the Uncertainty Principle, why this is not possible.

A pragmatic anti-reductionist can also argue that just knowing the properties of the parts is not enough to derive higher level properties; we also have to know the composition of the higher level entities that we are interested in. Thus although the set of valid higher level descriptions may be *determined* by the lower level properties, they cannot be *discovered* or *derived* without additional knowledge. Thus we find that, with a few exceptions in astrophysics, there are virtually no cases in the history of science in which a higher level scientific law or description has been derived from a lower one; rather such phenomena are discovered by investigation at the appropriate level and only subsequently related to lower level properties.

For a non-realist — for whom properties *only* exist within, or with respect to, our knowledge — these pragmatic objections to reductionism are also ontological ones. Thus if objects and properties must be discovered at their appropriate level (rather than the higher being derived from the lower) then they will have the same epistemological status; and this means that there will also be an ontological symmetry between levels of organisation. Indeed it seems plausible that it is this ability to avoid reductionism that attracts many materialists to some form of pragmatism or instrumentalism in the first place: instrumentalism allows one to be a materialist without implying reductionism.

(It is sometimes said that an abstract philosopher, in contrast to a hard-headed scientist, is one who will complain that 'it may work in practice, but does it work in principle?' We *know* that anti-reductionism is a vital strategy in scientific practice, so why should we care whether it works in principle? The problem is that there is no Chinese Wall between practice and principle. A scientist strives to get their truth-in-practice as close to the truth-in-principle as possible; their assumptions

of principle guide empirical work. Therefore if we want to avoid a reductionist scientific practice then we need anti-reductionist principles.)

Another way of avoiding the implications of reductionism is to express the relationship between levels of organisation in terms of *supervenience* (Kim, 1984). The idea is that a property of a whole,  $P$ , is supervenient on some properties of the parts,  $p$ , iff there can be no change in  $P$  without a change in  $p$ ; or if when two entities are indiscernible with respect to  $p$  they are indiscernible with respect to  $P$ <sup>1</sup>. Thus the concept of supervenience can be used to describe the relationship between levels of organisation without mentioning the reductionist bogey-word ‘determination’.

The notion of *emergence* has also been used to do a similar job to that of supervenience. The idea is that higher levels of organisation ‘emerge’ out of the lower, rather than being determined by it. But the problem is then to define emergence in a way that does not involve determination. One way is a kind of mystical holism in which wholes are blessed with properties that are not dependent on parts. But the more common way is a kind of pragmatic reductionism in which properties of wholes are considered emergent if they are in some way novel or surprising (Nagel, 1961, p374-80)(Crutchfield, 1994). Hydrogen and oxygen gas are not wet, for example. Indeed there is nothing about them which even *suggests* wetness. But if you put them together and spark a chemical reaction then you get water, which quite clearly is. Nonetheless it still seems the case that the properties of water are determined by those of its constituent molecules, even if we have problems deriving them.

Another way of combining reductionism and pluralism is to hold that each doctrine is true of its own separate domain. The usual split is that reductionism holds for biology downwards, whereas pluralism applies to humans and their cultures ‘from the neck up’, as it were. Thus we find that many philosophers are relatively uncritical of reductionism in natural science whilst strenuously denying that the same methodology can be applied to human affairs<sup>2</sup>. Nonetheless it is hard to avoid the blunt fact that our human experience is in some way linked to our ‘lower’ properties. If you push me, for example, then, as a physical object, I will fall. If I fall then, as a biological object, I will be injured. If I am injured then, as a psychological object, I will be in pain. And if I am in pain then, as a social object, I will sue. Thus we have many ways of being, and all these ways of being — these levels of organisation — are linked.

There is evidently some connection between levels of organisation, but what? The obvious answer is the reductionist one, that higher properties are dependent on lower ones. But in the rest of this chapter I argue that, although reductionism may be true in one sense, it only gives us half the picture. Reductionism is true in the sense that the properties of an object are indeed dependent on those of its parts; but it is *also* true that the properties of parts are dependent on wholes. Consider this example. In the nineteenth century Britain twice went to war with Manchu China in order to free up her trade in opium. So the foreign policy of the British Empire resulted in an increase in the concentration of opiates in the brains of millions of Chinese peasants. If a Chinese neurologist wanted to know why there was such a high concentration of endorphins in the

<sup>1</sup>These two versions of supervenience are not strictly equivalent, but the differences between them are not important for this discussion.

<sup>2</sup>The same combination of intuitions recurs in Margaret Thatcher’s infamous claim that ‘there is no such thing as society, just individuals and their families’. After all she did not claim that there are just *atoms* and their interactions, as a more consistent reductionist would. Individuals are the only things that exist for Thatcher for the bluntly pluralist reason that they are the actors in the political discourse that she was concerned with.

synapses of the brain cells she studied, then part of the answer would point to British Imperialism. Moreover, one of the effects of this mass addiction was to increase Chinese support for the Taiping rebellion which called for the prohibition of opium and the expulsion of the foreign powers that supported the trade. So in order to understand this piece of history we must trace the connections from the social level, down to biology, and back up to the social again; and it is precisely this kind of analysis that both reductionism and pluralism rule out. The pluralist would argue that you can understand the ebb and flow of historical tides without bothering with biology. The reductionist would disagree, but would argue that the biologist should stick to biochemistry, because politics cannot tell us anything about brains.

Or consider this other example. We are now all familiar with the increasing power and sophistication of psychoactive drugs that are able to relieve the symptoms of various conditions, such as the effects of SSRIs on depression. The reductionist interpretation of this success is that the psychological depression is *caused* by a neurochemical imbalance which the drugs correct. Many pluralists are unhappy with this interpretation and prefer to emphasise the psychological and/or social causes of the condition, but seem to believe that this requires they deny that the drugs have any beneficial effects at all. But there is no contradiction between the two explanations. It is perfectly possible for social pressures to have effects on the neurochemistry of our brains, with depression being the result. Drugs can break one link in this chain — and the relief can be welcome — but this does not imply that the condition was ultimately biological.

We are not determined by our biology, as the more ‘greedy’ reductionist would argue. But nor are we independent of it, as the more ‘idealistic’ pluralist would argue. What we do as humans depends on our biology; but it is equally true that what we do as humans *affects* our biology. The rest of this chapter is an attempt to outline an alternative way of understanding the relationship between levels of organisation that can accommodate this simple intuition.

## 2.1 Reductionism and Materialism

The basic premise of reductionism is that the world is made of objects, each of which has properties. These objects come together to form larger objects, and the properties of these wholes are dependent on the properties of their parts. We may then argue whether or not descriptions involving those larger objects are eliminable, or whether the properties of the wholes are strictly derivable from those of the parts, but the basic logic seems to be an irrefutable and inevitable consequence of materialism (Melnyk, 1995)— hence the suspicion in some quarters that anyone who espouses any form of anti-reductionist holism must be some kind of ‘flaky’ anti-materialist. Let me put the same point another way: how, precisely, can the whole be greater than the sum of the parts? If one is to remain a consistent materialist then it is a bit of the problem to explain where the extra comes from.<sup>3</sup>

Reductionism seems like one of the most obvious and basic truths in science. But it is precisely

---

<sup>3</sup>“The whole is greater than the sum of the parts” is a useful way of summing up the basic intuition of anti-reductionism, but it is strictly inaccurate. The idea that the whole is greater than a linear sum of its parts is perfectly compatible with even the strictest reductionism. The gravitational force between two masses, for example, is equal to the *product* of the parts, and yet this example is a triumph of reductionist analysis rather than being any kind of threat to it. A more accurate way of expressing the problem is to ask how the whole can ever be more than a *function* of the parts.

because it seems so obvious and basic that I want to put it under suspicion. The aim is to show how it is a peculiarly one-sided way of looking at the world, and to suggest another perspective in an attempt to redress the balance. Unfortunately the reductionist intuition runs so deep that it is difficult to know where to start challenging it. It is tempting to start from metaphysical first principles — this is, after all, the philosopher's natural strategy — but I have found that presenting the argument in this way rarely convinces. It seems that the intuition is just too deep. The alternative strategy that I pursue here is much more pragmatic. I consider three simple, familiar, examples of systems made of many interacting parts, and show how the anti-reductionist perspective can do useful work in making sense of aspects of these systems that the reductionist perspective neglects. I hope that the very mundane familiarity of these examples will convince where metaphysical generalities would not: if the anti-reductionist perspective can do some useful work on such well-worn examples then perhaps it should be given a chance on the more obscure ones discussed later in the thesis?

The first example is the Boyle-Charles Gas Law, which states that temperature of a closed container of gas is inversely proportional to its temperature. In this case we have an object (the container) that has properties (temperature and pressure) that behave in a particular way (they vary inversely). Maxwell and Boltzmann, in a triumph of reductionist analysis, proved how the behaviour of the gas could be explained by the motion of the individual molecules that make it up: each of these molecules collide elastically with each other and the container walls, and as we heat the gas the velocity of the molecules increases and they exert an increased force on the container walls. Thus the temperature and pressure of a gas are determined by the motion of its molecules, and the gas laws governing the properties of the whole container are determined by the laws governing the behaviour of its parts.

The second example is Conway's Game of Life. Suppose we have a large grid of square cells, each of which can be 'on' or 'off'. The state of each cell at the next step in the life-cycle of the grid is determined by simple rules defined over the current state of the cell and those of its neighbours. Out of these simple rules emerge a rich 'eco-system' of higher order patterns that may glide across the grid, blink between two states, generate gliders, and so on<sup>4</sup>. The Game of Life is often used as an illustration of how systems of interacting parts obeying simple rules can produce novel and interesting behaviour. But nonetheless it is still the case that the appearance and behaviour of the objects in the system (i.e. the higher order patterns) is determined by the arrangement and properties of their parts. Gliders glide and blinkers blink because of the rules governing the cells. The Game of Life is an example of how the reductionist approach can make sense of the emergence of complexity, not a challenge to it (Faith, 1998).

The third example is a car engine. A car engine is made of many different parts — driveshafts, pistons, cam-belts, and so on — each of which are carefully engineered to have very precise properties. None of these parts produce any power on their own, but when they are put together in the right way then we have a complete engine that does. Power is thus a property of the whole object that is dependent on the properties of the various parts.

It should be noted at this point that none of these three examples are biological or social. I do not, for example, consider how thought processes can emerge out of the interacting neurons in

---

<sup>4</sup>For more on the ontological and epistemological status of these patterns see Dennett (1991).

our brains, or how social systems can arise out of the interactions of free agents. This omission is deliberate for I want to break decisively with the intuition, mentioned above, that reductionism may be true of the lower sciences, but not for us higher, sophisticated, biological beings. This intuition only serves to enforce the gap between natural science and the philosophy of mind that it is my purpose to break down. I hope to show, by considering such ‘mechanical’ examples, that reductionism is not enough to understand examples from what is usually taken to be its strongest ground, and hence that its extension to other areas should be viewed with suspicion.

## 2.2 Anti-Reductive Materialism

So, how can an anti-reductionist perspective help us understand such simple systems? Consider the derivation of the gas laws from the kinetic theory of molecular collisions. Why do these derivations work so well? Feynman argues that

we shall find that we can derive all kinds of things — marvellous things — from the kinetic theory, and it is most interesting that we can apparently get so much from so little. . . . How do we get so much out? The answer is that we have been perpetually making a certain important assumption, which is that if a system is in thermal equilibrium at some temperature, it will also be in thermal equilibrium with *anything else* at the same temperature. (1963, p40-1)

So what happens if the gas is *not* in equilibrium?<sup>5</sup> The easiest way to find out is to compress it. As soon as we push on the walls of the container the measured pressure will rise, and as we continue to push we do work in compressing the gas. This energy diffuses through the container, raising the mean molecular momentum per unit volume, and those molecules nearer the compressed surface will be affected before those further away. Thus the properties of the parts are affected by what happens to the whole. The constituent molecules have the momentum that they do *because* of the pressure on the cylinder. The dependency only appears to run the other way when the system is static or in thermal equilibrium. Or suppose that we cool the container until the gas reaches its dew point where the molecules stop rebounding and start to stick together as the gas condenses into a liquid. Thus the molecules only collide elastically *because* they are in a gas at a certain temperature. Changes to the whole can affect the rules governing the behaviour of the parts.

The molecules of the gas are causally affected by what happens to the whole container, but there is another way in which the properties of parts of a system are dependent on the whole. This is *conceptual* dependence. Suppose, for example, we wanted to know the power of the car engine, and in order to measure this property we connected a measuring device to the main drive shaft. Now the drive shaft is clearly a *part* of the engine, and yet the power we measure is described as a property of the *whole*. We would not usually say that the power of the *shaft* was *X* Watts, but that the power of the *engine*, at the shaft, was *X*. Power is a property *of* the whole engine, but is located *at* one of the parts. On the other hand, when we measure the power at, say, the camshaft we would not normally describe this as ‘the power of the engine measured at the camshaft’. Why the difference in the two cases? What makes one a property of the whole and the other a property

---

<sup>5</sup>Non-equilibrium systems have been largely neglected in physics, with the notable exception of the work of Prigogine (1962).

of a part? The reason depends on the relationship between the engine and the rest of the car — i.e. the larger object of which the engine is itself a part. Now the output of the engine is connected to the rest of the car *via* the drive shaft, so the properties of the cam shaft and piston heads and all the other parts do not effect the rest of the car directly, but only through that single output. The power of the drive shaft is dependent on the properties of the other parts, not *vice versa*. In other words something is described as being a property ‘of the whole’ *because* it is dependent on other parts. Therefore the dependence of wholes on parts is built into the way we define ‘property of whole’ and ‘property of part’; it is a conceptual assumption that we make, not an empirical result about the way systems actually work.

The distinction between ‘higher’ and ‘lower’ properties in such cases is purely epistemic. It is an artefact of how we view the system. The speed of the camshaft, or the temperature inside the cylinders, is no less a property of the whole system than the power of the drive shaft. Each of these properties is at the same level. The distinction is rather between properties whose direct effects are felt outside the system and those whose effects are internal. We designate the former as higher properties simply because we cannot see under the skin of the system, when in strictly ontological terms there is nothing to choose between them.<sup>6</sup>

What about our third example, the Game of Life? The problem with the usual reductionist picture is that it treats the Game as a formal system. However all actual Games of Life — as opposed to the Platonic Ideal of the formal definition of the game — exist on computers<sup>7</sup>. And all computers — as opposed to the Platonic Ideal of formally-defined Turing Machines — exist in a physical and social context. They have power supplies, human users and programmers, cooling fans, manufacturers, and so on. This context forms a larger system of which the computer, along with the Game of Life that runs on it, is just a part. Moreover this context can causally effect the running of the game: the power supply may fail and interrupt it; the user may get bored and switch it off; the programmer may start to hack at the code; or the manufacturer may force an upgrade of the operating system which renders the old code obsolete. Therefore the behaviour of the gliders and blinkers in the Game are *not* determined solely by the rules governing the parts, but are also dependent on the physical context in which the Game runs.

However it may still be argued that in all cases in which the Game is running then the supervenience of the higher patterns on the cellular rules is maintained; i.e. as long as the Game is running ‘normally’ then the properties of the gliders and blinkers are determined by the rules governing the cells, even though the Game as a whole is dependent on what happens in its environment. But this is not quite true. Suppose we ask why the cells obey the rules that they do? The simple answer is that the computer was programmed in a certain way such that the cells obey the rules defined by Conway. But why was the computer configured in that way rather than any other? Now the rules of the Game were not revealed to Conway from on high, but were the result of experimentation; he tried many different versions until he found a set of rules that generated interesting behaviour. This process of experimentation is still going on. Most copies of the Game available on the Web, for example, allow the user to play with the rules themselves, and so there are many different versions

---

<sup>6</sup>This point will become important when we consider the distinction between theoretical and observational terms in general (section 3.3), and the status of mental representations in particular (5.2).

<sup>7</sup>It is also possible to run the game using pen and paper though this is time-consuming and tends to rob the game of its interest. The same arguments apply in either case, so I will just discuss the computer-based form.

of ‘the’ Game in existence. Thus, in all existing instantiations of the Game, the rules governing the interactions of the cells have the form that they do *because* they generate interesting higher patterns, as much as *vice versa*. If the rules did not generate interesting patterns then they would be changed.

(Of course within the space of all possible cellular automata there exists one, call it *L*, which has the same rules as Conway’s Game; and the rules of *L* are prior to, and not dependent on, the behaviour they generate. But unless we invoke the axiom of choice then *L* is picked out as The Game *because* of the higher behaviour. Therefore, even within the space of possible CAs, the property of ‘being the rules of the Game of Life’ is not prior to the property of ‘being the emergent patterns of the Game of Life’. The rules may be defined independently of their emergent behaviour but this does not imply that they are ontologically prior.)

Systems like the Game of Life certainly exhibit rich and fascinating behaviour. And in some cases such systems can be successfully used to model natural biological phenomena — such as in the work of Thom, Waddington, Kauffman, and Goodwin, and in Turing’s diffusion-reaction model of morphogenesis. But we should be careful about deriving general philosophical conclusions about the relationship between levels of organisation in nature from such artificial systems. These models embody certain assumptions about how physical systems work. In particular, they assume that there is a set of prior lower level entities whose behaviour is determined by fixed laws. Therefore, when we find that their higher level behaviour is only non-reducible in a weak, pragmatic, sense we should not assume that this is a correct understanding of emergent phenomena in nature. Dennett once noted that, for philosophers, the attraction of experiments such as the Game of Life is that one gets to make up the facts. But we should be aware of the cost of such factual liberalism.

Where did the reductionist go wrong? Where is the flaw in their argument about the inevitability of reductionism? The problem was that the reductionist starts by considering the properties of objects in isolation, and then asks what happens when those objects come together to form wholes. The reductionist metaphysical intuition is that objects are in a strong sense *independent* of their environment, in the sense that they need nothing else in order to be. According to this intuition properties are *intrinsic* to objects, they are *essential*, they belong to the *object-in-itself*. But no object has ever existed ‘in itself’. All objects exist *in the world*. Although we can imagine objects on their own — we can leave the mental background blank, as it were — all objects that have ever actually existed have done so in environments. All things are, on every occasion, surrounded by other things. All objects are parts of larger wholes. Molecules are parts of gasses, engines are parts of cars, and Games of Life exist within computers. The same is true of stars in galaxies, individuals in societies, cells in bodies, neurons in brains, and right down the tertiary structure of proteins in their enzymatic environment. In none of these cases are objects born in isolation and subsequently come together to form wholes, rather the object comes into being as part of the whole.

Moreover the properties and rules governing the behaviour of objects *depend* on the properties of those larger wholes, and they do so in two ways. First, properties of parts are *causally* dependent on the properties of wholes: cooling the container causes the molecules to stop colliding, and the gliders and blinkers in the Game of Life stop gliding and blinking if the program is interrupted.

Second, properties of parts are *conceptually* dependent on wholes: the power measured at the drive shaft ‘is’ the power of the engine because of the way in which it is connected to the rest of the car, and the rules of a particular CA ‘are’ the rules of the Game of Life because of the types of pattern that they generate<sup>8</sup>. Wholes are often described as being emergent products of their parts. This is true, but it overlooks the fact that *parts* are also emergent products of *wholes*.

### 2.3 Downwards Causation

Downwards causation (i.e. the causal dependence of parts on wholes) has had a disreputable history in the philosophy of science ever since it was proposed as a solution to the mind-body problem by Sperry, Popper and Eccles (1977). It has also become unpopular in the context of social studies by its association with strongly structuralist analyses of history, in which the actions of the individual are determined by higher social structures. Szentágothai admits that defending downwards causation will confirm his image as a ‘crazy Hungarian and an impossible romantic adventurer’ (1984), whilst Bedau notes that

although [downward causation] is logically possible, it is uncomfortably like magic. How does an irreducible but supervenient downward causal power arise, since by definition it cannot be due to the aggregation of the micro-level potentialities? Such causal powers would be quite unlike anything within our scientific ken. This not only indicates how they will discomfort reasonable forms of materialism. Their mysteriousness will only heighten the traditional worry that emergence entails illegitimately getting something from nothing. . . . But the most disappointing aspect of [downwards causation] is its apparent scientific irrelevance. . . . We should avoid proliferating mysteries beyond necessity. To judge from the available evidence, [downward causation] is one mystery which we don’t need. (Bedau, 1997, p377)

But downward causation is neither mysterious nor superfluous. For example, suppose we measure the mean momentum of a particular gas molecule over a period of time, and then compress the container by 20%. From just this information we can accurately predict that the mean momentum of the molecule will rise by a proportionate amount. This is an example of prediction using downwards causation that is both easy and reliable. The reductionist would claim that we could have produced a similar prediction given enough information about the exact trajectories of the other molecules. But this ignores the fact that molecules are not perfectly elastic billiard balls. If they were then gases would never condense. The determined reductionist would also need information about the exact structure of the electronic orbits of the molecules in order to predict the results of their interactions. In contrast the analysis that uses downwards causation is easy, reliable, and theoretically sound. If one ignores the kind of empirical regularity on which it is based then one has missed an important fact about the behaviour of the system. In short, if one is inclined towards pragmatism — with a small ‘p’ — in science, then downwards causation is as pragmatically useful and theoretically respectable as any other sort.

The same argument can also be cashed out in terms of counterfactuals (Lewis, 1986). Suppose one holds that *C* causes *E* iff *E* would not have happened if *C* had not. In our example the stochastic

---

<sup>8</sup>Conceptual dependence will turn out to be important when considering the relationship between mental states and the world (section 5.3), and between genes and organisms (9.3). But for most of the rest of this chapter I concentrate on the strictly causal relationship between parts and wholes.

nature of the system ensures that the effect of compressing the gas on the average momentum of the individual molecule would have been the same no matter what particular sequence of collisions occurred: the lower level events are overdetermined by their immediate causes. Therefore it is the higher, rather than lower, events that should properly be described as the cause in this case.

One objection to the possibility of downward causation is that it seems to imply causal overdetermination: if the momentum of a molecule is caused by its own history of collisions then how can it also be caused by the gas being compressed in a cylinder? But the idea that there must be a unique efficient cause of any event is an unnecessary hangover from Aristotle. Any real phenomenon is a dense web of causal processes. We can usefully pick out certain of these processes as being more important for statistical or discursive reasons, but this does not require that those properties be regarded as the *unique* cause of an event. Of course all instances of downwards causation will be mediated by local, lower, mechanisms. But once we abandon Aristotle's insistence on a unique efficient cause, then there need be no contradiction between regarding both higher and lower events as proper causes.

However there is one notion of cause that does necessarily exclude the possibility of downwards causation. This is to demand that observed correlations are only causal if they are instances of a covering law. So, for example, compressing the cylinder may effect the motion of particular molecules, but the laws governing their motion will remain the same. According to this view it is the frequency and order of micro-events which are the target of downwards causation, rather than the laws that govern them (Schröder, 1998). This version of anti-reductionism concedes that objects may be affected by their environment, but the *way* they are affected is a fixed and intrinsic property of the object. For example, moving the piston of a compressed gas will change the motion of the constituent molecules, but the *laws* that govern their collisions remain unaffected. According to this analysis downwards causation can always be eliminated — at least in principle — in favour of a causal story written in exclusively lower-level terms. The causally effective higher level property is just a shorthand description of a state properly defined at a lower level. Thus we can re-write the claim that 'the movement of the piston affected a molecule' in terms of the collision of individual particles. But as we saw above, gas molecules do *not* always obey the laws of elastic collisions. If we cool the container then the molecules stop rebounding and start to stick. Therefore changes in the environment of a part do not just affect the part. They can also change the *way* the part is affected by its environment.

There are two ways of dealing with such examples. The reductionist reaction is that if a property turns out to be dependent on the context then it should be eliminated in favour of one at a lower level. So if molecules are going to stop rebounding and start sticking then perhaps the gas should be understood at the lower, and presumably surer, level of atoms and electronic orbits. The reductionist axiom that objects and their properties are in a strong sense independent of their environment is built into their definition of a 'real object' and 'real property'. Reductionism is an *a priori* assumption about how the world is: if properties and entities are dependent on their environment then their reality is questioned; they are second-rate entities, just approximations that we find convenient. But it is highly doubtful that the reductionist strategy would *ever* yield properties that satisfy their criteria. The reductionist ends up in free fall, tumbling through the levels of description, looking for one that fits the ideal of a billiard ball universe. Until the revolution

of quantum mechanics there seemed to be a bottom level safety net, but now even that has disappeared. However the philosophical problems around quantum mechanics are eventually resolved — whether by the use of action at a distance, or the role of the conscious observer in the collapse of the wave-function, or some other equally exotic solution — it seems more than likely that it will involve some form of radical environmental-dependence of properties and fail to fit the classical reductionist picture. Quarks, for example, are currently assumed to be the ultimate components of nature, and yet they cannot even exist outside of the protons and neutrons they form. The reductionist criterion of objecthood is that an object does not require anything else in order to exist, but quarks do not meet this criterion. The persistent reductionist is in danger of being unable to find anything to reduce higher level descriptions to<sup>9</sup>.

The alternative strategy, the *anti-reductive* strategy, is to accept that properties are held by objects-in-environments, not objects-in-themselves. There are no such thing as truly intrinsic (i.e. environment-independent) properties since nothing ever exists in isolation. The situations that science usually describes as ‘isolated’ — such as a vacuum with flat electromagnetic and gravitational fields, or a laboratory arena, or a test tube or Petri dish — are themselves environments as much as occupied space, natural environments, or living organisms are; they are just different kinds of environment. (Similar considerations force us to conclude that one does not find one’s ‘true’ self in a Buddhist retreat, just a different one.) When we say that an object weighs *X*, or has a mass of *Y*, or that it has the colour *Z* — i.e. whenever we predicate a property of an object — we actually mean that the object *on earth* weighs *X*, or has a mass of *Y in our inertial frame of reference*, or has colour *Z at room temperature*. If we put the object on the moon, or in a rocket, or in a kiln, then its properties will change accordingly. Objects only ever exist in environments, and the environment can affect even the seemingly most fundamental and intrinsic of properties. An object may be made of gold, but put it in *aqua regia* and it soon dissolves; therefore its continued existence as a gold object is dependent on its environment *not being aqua regia*.

Science proceeds by trying to find objects and properties that are ‘robust’ — i.e. that are constant across environments — and it is often successful. But we can draw two different conclusions from this success. The first is that those properties are essential and intrinsic to the object, and held independently of the environment. Or we can conclude that those properties are held in, and due to, that range of environments. Although the two positions account for any given set of empirical evidence equally well, the former position is far stronger in that it implies (or, rather, assumes) that the same properties will be held in other, future, environments. Science is built on extrapolating from observed cases to unobserved situations. We assume that the physical constants measured in the accelerator are the same outside our light cone, that the biochemistry of the Petri dish proceeds in the same way in the living cell, or that the psychological behaviour exhibited in the laboratory would happen in everyday life. Sometimes this is true but, as Feynman points out, in each case this is an assumption and should be acknowledged as such:

Of course this means that science is uncertain; the moment that you make a proposition about a region of experience that you have not directly seen then you must be uncertain. But we must make statements about the regions that we have not seen, or the whole business is no use . . . We have to make guesses in order to give any utility

---

<sup>9</sup>B.C. Smith (1996, ch5) also argues that modern field-theoretic physics does not supply us with ready-made objects in the way the reductionist fondly imagines it does.

at all to science. In order to avoid simply describing experiments that have been done, we have to propose laws beyond their observed range. There is nothing wrong with that, despite the fact that it makes science uncertain. If you thought that science was certain — well, that is just an error on your part. (1965, p76)

Or, as Hume put it 200 years before,

Even after the observation of the frequent conjunction of objects, we have no reason to draw any inference concerning any object beyond those of which we have had experience. (*A Treatise of Human Nature* II.3.12)

These are simple truths, but they are ones that the reductionist seems unwilling to acknowledge.

## 2.4 Conclusion

The debate between reductionism and anti-reductionism is a debate over what kinds of questions one should ask in order to understand how physical systems work, and the kinds of answers one should expect. Now the reductionist certainly asks valid questions, and their answers are certainly true and useful. But we should not infer from this success that their answers are complete. In particular they ignore the way in which objects are dependent on their surroundings, and not just their insides. This simple intuition will turn out to have important consequences, especially when the objects under consideration include intelligent agents.

## Chapter 3

### Naturalisation

---

From Man or Angel the great Architect  
Did wise to conceal, and not divulge  
His secrets to be scann'd by them who ought  
Rather admire . . .  
Solicit not thy thoughts with matters hid,  
Leave them to God, Him serve and fear.  
— Milton, *Paradise Lost*

‘What is *internal* is hidden from us.’ — The future is hidden from us. But does the astronomer think like this when he calculates an eclipse of the sun?  
— Wittgenstein, *Philosophical Investigations*

In 1609 Johannes Kepler published a book, *Astronomia Nova* (The New Astronomy), in which he proposed two laws that described the motion of the planets in terms of ellipses focussed on the sun. This is rightly seen as one of the great defining achievements of science, indeed as one of the great achievements of *humanity*. Why? What is it about Kepler’s discovery that epitomises our ‘idea of the good’ in science?

Kepler’s genius lay in combining an old but controversial idea with a new, even more controversial, idea of his own. The old idea was that the planets went round the sun. This had first been proposed by Aristarchus of Alexandria, but unfortunately he also assumed that the motion of the planets must be circular — and Hipparchus later showed that this combination did not fit astronomical observations. Hipparchus dropped Aristarchus’ heliocentricity but retained the assumption of circularity (since this was obviously the most ‘natural’ type of path), and from this Ptolemy developed a terracentric astronomy based on epicycles. If enough circles were stacked upon each other then the terracentric astronomy could be salvaged. It was messy — 77 epicycles were eventually needed — but it worked. In 1543 Copernicus showed that the epicyclic system could be greatly simplified if Aristarchus’ proposal of heliocentricity was resurrected. The 77 epicycles could be reduced to 31, and even greater accuracy could be achieved by shifting the sun slightly off-center. In mathematical terms Kepler’s great achievement was to show that the heliocentric system could be simplified still further by replacing the epicycles with ellipses, with the sun at one focus rather than at the geometric center. All the known astronomical data could

then be accounted for using just three<sup>1</sup> simple laws.

But this was not, in itself, enough to secure Kepler's place in posterity. The history of science is usually written as 'Whig' history — i.e. from the point of view of the winners of scientific disputes — and it is sometimes forgotten how Kepler's theory was derided at the time. Its empirical accuracy was disputed by no less a personage than Francis Bacon, father of empirical science. The brilliant Cardan refuted its mathematical basis. Kepler and Copernicus were lambasted by Martin Luther, Calvin, Montaigne, and Milton, and satirised by Ben Jonson and Shakespeare. The Roman Inquisition persecuted Galileo for even suggesting that their ideas may have some merit.

It was Newton, working in England beyond Rome's grasp, who saved Kepler. In 1687 Newton published the *Principia*, which showed how the elliptical motion of the planets could be *explained* by the force of the sun's gravity. Kepler had always intuited that there must be some heliocentric force that kept the planets in their elliptical orbits (Stephenson, 1997), and Newton demonstrated that this force was the same as that which could be directly observed acting on apples here on earth. But the story did not end there. According to Kepler's theory the axes of the planetary orbits are fixed. But in the 18th Century it became clear that the perihelion of Mercury was slowly advancing. Leverrier showed that part of this shift could be explained by the gravitational effects of other planets, but a significant part remained a mystery. At the start of this century Einstein rewrote the Newtonian Book, and in doing so explained the discrepancies in Mercury's orbit. But notice this. Ptolemy and Copernicus' contributions to physics were effectively deleted when Kepler added his new chapter to the Book of Physics. The theory of epicycles is now only of historical interest. But when Einstein added a chapter, Kepler did not join Ptolemy in the dustbin of scientific history. His theory remains a vital brick in our understanding of the physical world. It is still, in a sense, *True* — despite Einstein. What qualities does Kepler's theory have that have made it so robust? Why does it seem insulated against possible refutation? Why is Kepler seemingly immortal? In this chapter I try to outline the sense of scientific truth that Kepler's — and other similarly immortal theories — embody.

### 3.1 Descriptions and Biases

Science proceeds by collecting empirical data and then trying to find patterns in it. The pattern is a way of describing, of making sense of, the data; and these descriptions are the basis of our theories. Of course the experimental scientist usually has an intuition of what patterns they are trying to find, and for them the key problem is creating an experiment in which the patterns show up in the data. Nonetheless they still have to make that crucial step from data to pattern. The problem is that every set of data contains a myriad different patterns. The same data can be described in many different ways. Tycho Brahe's astronomical data could be described in terms of Kepler's ellipses or Copernicus' epicycles, so how should we choose between the various possible theories? (I use the terms *description* and *theory* interchangeably, since every particular description falls under a general theory, and our theory informs our choice of description.) Rorty describes this process of choosing a way of describing empirical data as 'adopting an attitude', Dennett describes it as 'adopting a stance', and in the field of machine learning it is known as a 'bias'. I will use the latter term because I want to avoid the some of the associations that Dennett and Rorty have drawn from

---

<sup>1</sup>The third was added ten years after *Astronomia Nova* in *Marmonices Mundi* (Harmony of the World).

theirs, though the intent is roughly the same.

What kinds of descriptive biases are there? In the everyday practice of both scientists and lay persons the main descriptive bias is social: we describe phenomena in certain ways because that is how we have been brought up and trained to do so. But how do we know that this is the best way? Surely we need some criterion for evaluating our current practice? On the other hand, the naive realist argues that our bias should be for the truth, that we should describe things as they really are; but how do we know what the truth is? The poet's bias is to describe phenomena in the way that best communicates her subjective experience to others, but the purpose of the poet is different to that of a scientist. (Melville, for example, devotes an entire chapter of *Moby Dick* to explaining why Ahab's whale was best described as white, even though a naturalist may insist that it was 'really' a dirty grey.)

Ockham and Mach, in their own ways, argued that the best theory is the simplest, all other things being equal. However this is an *a priori* bias: of course it is nice if things turn out simple, but it hardly seems justified to impose our tastes on nature. Moreover, we can always increase the simplicity of a description by reducing its accuracy and disregarding some of the data as noise. Simplicity and accuracy therefore form two conflicting biases and we need some way of arbitrating between them in order to separate noise from the 'real' data. Rutherford, for example, discovered the atomic nucleus by bombarding gold foil with  $\alpha$ -particles. Most were deflected slightly as they passed through the electronic cloud of a gold atom in the foil, but a very few rebounded as they hit the tiny nuclei directly — like 'cannon-balls bouncing off a sheet of tissue paper'. The simplest, and overwhelmingly accurate, description of this data would have been to disregard the rebounds as noise and just account for the partial deflections using a model of continuous charge distribution. Therefore Rutherford had to use biases *other* than simplicity to justify his description of the phenomenon. The simplest theory may be the best, all other things being equal; but what other things?

The most popular bias in the philosophy of science is that the best theory is the one that yields the most accurate predictions. (It is also the bias that most philosophically-minded scientists would claim that they adhere to.) Now it is certainly true that one of the purposes of science is to predict the future<sup>2</sup>. But there is something distinctly odd about this stance: why should facts about the present depend on the future? The thing we are trying to describe has already occurred and now persists in our recorded observations, and yet the predictivist claims that the way it should be described depends on future events. This only makes sense to the extent that we assume that there is a fundamental constancy — a lawful regularity — in the pattern that we are describing that has existed up to now and will persist into the future. Not so much '*que sera sera*' as '*whatever has been will be*'. If this is the case then those future events will shed further light on the nature of the pattern already observed, and the failure of a description to predict the future is a good sign that it has failed to capture something about the present.

(This problem of future-dependency is not just philosophical, but also methodological. Empirical scientists, such as meteorologists or geologists, who create mathematical models of complex systems face the practical problem of how to choose between competing models. For such scientists the philosophical principle that the best description is the most predictive is not much

---

<sup>2</sup>The social origins of the bias of prediction will be discussed in 11.2.

methodological help: they must choose now, on the basis of the available data, which model to accept. Productiveness may be a good way to judge descriptions retrospectively, but is not much help in forming them, as Oreskes notes (1994.)

The predictivist could respond to the problem of future-dependency by arguing that whether or not a description will prove to be the most predictive is a fact about the current state of the system, even though we cannot use this to describe a system without observing its future behaviour. Therefore the future-dependence of the description is only epistemological rather than metaphysical. But the future behaviour of the system — and hence the correct description — is not necessarily fixed by the data that we are trying to describe. The reason for this was first pointed out by Babbage (1864): for any given system and observed behaviour we can construct another system that displays the same behaviour up to a given time,  $t$ , but subsequently diverges. Therefore even though the behaviours of the two systems up to time  $t$  are identical, the correct descriptions of them are not. The correct description of a system is *not* determined by the behaviour we have observed up till now.

For example, suppose we observe two pool-players. The first is not very good and gets easily beaten. The second *appears* to be not very good but she is in fact a hustler, and as soon as she has persuaded an opponent to put some money down then she raises her game. Until there is money on the table the behaviour of the two players appears to be identical, but the correct description is not: one is playing pool badly, and the other is losing on purpose. But if we cannot see the future, then how can we choose between the two descriptions? If the correct description of the behaviour of a system is not determined by its observed behaviour then what else could it depend on? In the rest of this chapter I discuss what that else could be — i.e. what bias we could use other than predictivity — and draw out some implications for our understanding of scientific explanations. However in this chapter I will *not* give a reason for preferring this alternative bias to that of predictivity. That will have to wait to the end of the thesis.

### 3.2 Naturalisation

The alternative bias to predictivity is this: in order to describe the behaviour of a system we cannot just rely on the observed data, we also have to *look inside* the system and understand how it works. Consider this toy example, introduced by Sober (1982):

Imagine a machine that sorts out wire shapes. It is made up of two components. The first operates as follows: when given a piece of wire as input it will output the wire if and only if the wire is a closed figure with straight sides. The second takes any number of straight pieces of wire and outputs them if and only if they have three angles; thus it will allow through an open four sided figure, but not a closed one. Therefore only triangles will pass all the way through. The question is, how should we describe the behaviour of the machine? Does it detect *trilaterals* or does it detect *triangles*? Now at first glance it seems like the two descriptions are exactly equivalent. After all, all triangles are also trilaterals. Therefore if the machine is detecting triangles then, logically speaking, it is thereby also detecting trilaterals. *And* the two descriptions will be equally predictive: if you show me a shape then I will be able to predict whether it will pass through the machine using either description. However once we understand how the machine works we can see that it was the number of angles in the closed figure that mattered, *not* the number of sides. What

the machine *does* — as opposed to what its behaviour *is* — is to detect triangles, not trilaterals. So once we understand how a system works — i.e. the mechanism underlying its behaviour — then we can use this information to choose between two equally accurate, and predictive, descriptions.

Let us call this bias ‘naturalisation’. When we understand how something works we make its behaviour non-mysterious, we make it seem *natural*, the behaviour becomes of its *nature*. It becomes clear why things of that type behave in that way. It was this process of naturalisation that saved Kepler. If you are only interested in empirical accuracy or prediction then, given enough epicycles, the Copernican (or even Ptolemaic) description of the solar system can be made just as accurate and predictive as one based on ellipses. (Indeed the Mayans were able to predict eclipses and the positions of the moon and Venus very accurately just using arithmetic and without invoking the notions of ‘orbit’ or ‘planet’ at all.) But Newton showed that only elliptical motion could be *explained* by a heliocentric gravitational force.

Darwin’s theory of natural selection is another paradigm case of the importance of naturalisation. Darwin was not the first to propose that species evolved. But until then evolution was regarded in England largely as the ideology of non-conformists, socialists, and continentals (and at the time it was hard to decide which was worse). Nor was Darwin the first to argue that organisms are adapted to their environment; but until then the only possible explanation for this had been God. Darwin’s achievement was to demonstrate the mechanism underlying evolution — descent with modification — which proved that it was in the nature of species to incrementally evolve and adapt, rather than be fixed, perfect, types. Moreover Darwin’s theory — like Kepler’s — was initially treated with scepticism. Mendel saved Darwin — just as Newton saved Kepler — by uncovering the mechanism underlying descent with modification.<sup>3</sup>

(Once Darwin had demonstrated how species become adapted to their environment, then a new kind of explanation — and a new way of describing the world — became scientifically respectable. This newly-respectable way was *functional explanation*, i.e. explaining the behaviour of a system by the way it fits into a larger whole, rather than its underlying mechanism. This is an example of explanation from above, rather than below. But, as we shall see in chapters 7 and 10, the validity of functional explanation rests on a Darwinian explanation of the mechanism through which the larger system evolves.)

All the great revolutions in science have involved realising that entities and behaviours which were previously thought to be fixed and ‘God-given’ were in fact inconstant: species, planetary orbits, inertial mass, gravitational mass, space-time, atomic nuclei, continents, aristocracies. However these revolutions did not replace an assumption of constancy with one of random change, but with a more precious ability to *explain* those changes through an understanding of the forces underlying the patterns that were previously thought to be constant. These revolutions went hand in hand with — sometimes preceding and sometimes following — new ways of describing the patterns observed in nature: natural selection rewrote taxonomy, planetary epicycles gave way to ellipses, energy and mass were equated, elements were ordered in the periodic table and further subdivided into isotopes, the old maps of land masses were ripped up in favour of ones based on tectonic plates, and Divine Right and the Three Estates gave way to the Rights of Man and the Social Contract.

---

<sup>3</sup>This will be discussed further in chapter 8.

The history of science is not just a steady accumulation of empirical data fitted with more and more accurate curves. It also involves transformations of our understanding of what we have already observed. Kuhn (1962) famously described these transformations as ‘paradigm shifts’ in which inconsistencies between data and theory reach a critical point and the way becomes clear for a new theory to be accepted. But the examples of transformations mentioned above can be better understood as successive *naturalisations* made possible by the discovery of the mechanisms underlying previously observed phenomena. The job of the scientist is not to simply collect data and then fit a curve; but is rather to use that data as a starting point for further investigation into how the observed system works: to turn the unobserved into the observed. Science is inherently progressive, not just in the quantitative, extensive, sense of being a steady aggregation of accumulated data, but also in the qualitative, intensive, sense of involving transformations in our understanding of what we have already observed. Our descriptive biases should reflect this progressive nature.

### 3.3 Theoretical Terms, Dispositions, and Causal Explanation

What, exactly, does naturalisation involve? I do not believe that it is possible to give a formal definition since, apart from anything else, scientific theories are rarely completely formal. (Mathematical physics is the exception rather than the rule in this respect.) Cussins, for example, argues that the *only* thing that defines a successful naturalisation (or ‘unification’, in his terminology) is that it makes the connection between the observed behaviour and underlying mechanism ‘intelligible’ (1992b). However we can pin down the notion of naturalisation more precisely in the way that it treats *theoretical terms*. These are terms used in our descriptions of the behaviour of a system that do not depend directly on observation, but are introduced in order to make sense of those observations. An elliptical orbit, for example, is a theoretical term. We never see an ellipse carved out in the sky. All that we observe directly are the positions of the planets at particular times, but in order to make sense of those observations Kepler introduced the notion of an elliptical orbit that the planet ‘follows’.

Reichenbach distinguished between two ways of regarding theoretical terms. *Illata* are entities that our observations suggest exist, whilst *abstracta* are logical constructs from observational terms:

Our observations of concrete things confer a certain probability on the existence of *illata* — nothing more. ... Second, there are inferences to *abstracta*. These inferences are ... equivalences, not probability inferences. Consequently, the existence of *abstracta* is reducible to the existence of *concreta*. There is, therefore, no problem of their objective existence; their status depends on a convention. (Reichenbach, 1938, p211-12)

Now how you regard theoretical terms depends on what you want out of your theory. Quine (1951), for example, requires only that theories should be predictive and concludes that the terms introduced by those theories are only ‘real’ to the extent that they help those predictions:

As an empiricist, I continue to think of the conceptual scheme of science as a tool, ultimately, for predicting future experience in the light of past experience. Physical objects are conceptually imported into the situation as convenient intermediaries — not by definition in terms of experience, but simply as irreducible posits comparable,

epistemologically, to the gods of Homer. Let me interject that for my part I do, *qua* lay physicist, believe in physical objects and not in Homer's gods; and I consider it a scientific error to believe otherwise. But in point of epistemological footing the physical objects and the gods differ only in degree and not in kind. Both sorts of entities enter our conception only as cultural posits. The myth of physical objects is epistemologically superior to most in that it has proved more efficacious than other myths as a device for working a manageable structure into the flux of experience.

Hempel regards theories in the same way as Quine, and introduced the analogy of a theory being like a net laid over the ground of our empirical experience (1965). The net is tied down at various knots, as certain terms of our theory are tied to observable data; but the other knots are not so fixed, connected to the ground only via a network of theoretical connections. In Hempel's picture there is no way to choose between two possible nets — and so two sets of theoretical terms — as long as they can both be tied to the same fixed observable points. Van Fraassen (1980) similarly insists that we should remain agnostic about the 'real' status of theoretical terms. On the other hand if you want to *naturalise* a theory, rather than just use it to make predictions, then your attitude to theoretical terms changes accordingly. Naturalisation requires that we make the observed behaviour non-mysterious. And if the theory that describes that behaviour invokes theoretical terms, then naturalisation requires that we make their ability to play a role in that theory non-mysterious. Planets, for example, follow elliptical orbits. Why? Kepler himself did not have an answer. For Kepler elliptical orbits were just the path that planets followed. They were abstracta. But Newton supplied an explanation of planetary motion by proving that elliptical orbits are minima in the energy field of the planet-sun system<sup>4</sup>. It takes an external force to shift a planet from this orbit, so an undisturbed planet will follow it in the same way that a marble follows a groove. Thus we have an understanding of what the theoretical term refers to *independently* of the behaviour that it is introduced to explain: elliptical orbits are grooves that planets follow, not just paths that they trace out. Newton turned Kepler's abstracta into illata.

Naturalisation requires that if we want to explain the behaviour of a system, *S*, by reference to its possessing a property labelled with the theoretical term *P*, then it must, at least in theory, be possible to determine whether or not *S* possesses *P* independently of the behaviour it was invoked to explain. If we cannot individuate the theoretical term in this way then it is no more than an empirically useful convention, rather than part of an explanatory understanding of the system. Sometimes we can observe *Ps* directly, such as when Crick and Watson discovered the mechanism underlying Mendel's genetic factors. In other cases — such as Newton's naturalisation of Kepler — we can only observe the forces out of which the theoretical term is constructed. No-one has seen an elliptical orbit, but we have all seen the effect of gravity from which those orbits can be calculated. But even in these cases the stuff out of which the theoretical term is constructed is not the same stuff that we are using that theoretical term to explain. The elliptical orbit of a planet is determined by its initial state and the sun's gravity, not the subsequent motion that we are using that orbit to explain.

The same argument applies to Dennett's example of a center of gravity (1991). Newton observed bodies acting under gravity and postulated a point through which the force acts. If we want

<sup>4</sup>Actually Newton did not conceive of orbits in quite this modern way, but Feynman shows how the two views are equivalent (1964).

to know *why* gravity acts like this it is not enough to explain that the center of gravity is the point through which gravity acts on a rigid body. It is empty, at best, to claim that gravity acts through a certain point *because* it is the center of gravity, unless we supplement this with an explanation of how and why centers of gravity have the properties that they do. The explanation goes like this. Newton's theory only describes the effect of gravity on point masses, but planets are large and complicated, made up of many parts that exert gravitational influence on each other. However if we assume that the body is rigid then Newton's law of action and reaction means that these internal forces cancel out, and that the net force on the whole will act through the weighted mean of the positions of the constituent point masses. This also explains why, when the body is not perfectly rigid, gravity does *not* act solely through the center of gravity (which is why we have two tides a day, rather than just one). We cannot observe centers of gravity directly, but we can observe the force of gravity acting on the *parts* of a large body, such as when we use a swing pendulum to measure the mass of a nearby mountain. From this evidence, and Newton's third law, we can explain how gravity will act on the whole.

Without a naturalised theoretical term we don't have an explanation, just a description. It was on this point that the wrong, but empirically successful, theories that litter the history of science tended to come unstuck. The caloric theory could account for the flow of heat, but it was molecular motion that could be observed buffeting Robert Brown's pollen grains. Epicycles and crystalline spheres could account for the planetary orbits, but only ellipses could be explained by a force that could also be observed acting on apples. Paley's God-designer could account for the origin of species, but only Darwin's descent with modification could be seen in the work of pigeon fanciers. Maxwell's equation could be understood in terms of aethereal vibrations, but only photons could produce the photoelectric effect. *Chi* is a very useful theoretical term in the hands of a Chinese doctor, but cannot be unified with the view through the microscope.

The problem of theoretical terms is closely related to the problem of dispositions. Carnap (1953) pointed out that if the dispositional property of 'being soluble' is defined as 'dissolving when in water' then the claim that 'X dissolved because it was soluble' is tautologous<sup>5</sup> But we can avoid this tautology if we regard a disposition as a kind of theoretical term that we invoke in order to explain the observed behaviour. And if dispositions are a type of theoretical term then the obvious next step is to turn them into *illata*; in others words identify a property of the substance *in virtue of which* it displays that behaviour. As Sober (1981, p149), following Quine (1969), puts it:

We characterise Quine's position that no irreducibly modal properties are permitted in science by saying that a property term which is defined counterfactually must be rendered epistemologically accessible. Although the predicate may be modally defined in terms of what would happen in some (nonactual) possible world, it should be possible to find out if the objects in the actual world possess that property.

For example, a substance will dissolve in water if the Van der Waals bonds between its molecules are weaker than the bonds that would be formed between those molecules and  $H_2O$ . This prop-

---

<sup>5</sup>This argument originates in Moliere's pastiche of 18th Century doctors who explained that opium induced drowsiness because it possessed 'dormative properties'. The same kind of doctors can be found today. Think of those who, for example, explain that a child is hyperactive because he has Attention Deficit Hyperactivity Disorder, when ADHD itself is defined in terms of the exhibited symptoms. ADHD does not explain hyperactivity, it just labels it.

erty is ‘epistemically accessible’ — we can find out whether a substance possesses this property *without* putting it in water. Therefore this is the property that the disposition of solubility refers to.

This argument for the naturalisation of theoretical terms and dispositions stem from the intuition that they should be regarded as causal properties and entities. Elliptical orbits, centers of gravity, solubility, and descent through modification, play a causal role in the phenomena that they were introduced to explain. If we do not require our theoretical terms to play a causal role, then there is no harm in leaving them as abstracta. The problem of causation is too deep to be tackled here<sup>6</sup>, though it is possible to say this: if we want to claim that an entity or property is a cause of an effect then, at the very least, it must have an identity independent of that which it has been invoked to explain. To say that ‘the cause of *A* caused *A*’ is empty, as Davidson (1980) puts it. If *A* is an observed behaviour, and the causes we are looking for include theoretical terms and dispositions, then naturalisation provides one way of individuating the cause of *A* independently of *A*.

The possibility of naturalisation is what differentiates physical objects from Homer’s gods. They differ in kind and not just degree, as Quine would have it. Gods are theoretical terms that we introduce to explain the world; likewise spirits and souls and *chi*. But what evidence do we have for them, other than that which we invoke them to explain? What are gods, or souls, made of? Why are they able to have the causal effects that we attribute to them? How do they *work*? We cannot ask these questions of gods, but they can be asked, and answered, of physical objects. In other words physical objects can be *naturalised*. But not all of the theoretical entities introduced by science live up to these requirements. Explanation must bottom out somewhere and so when it comes to the bottom level of explanation, to the most elementary physical particles, then naturalisation is not an option. As elementary particle physicist James Cushing remarks (1982, p78)<sup>7</sup>,

When one looks at the succession of blatantly *ad hoc* moves made in quantum field theory (negative-energy sea of electrons, discarding of infinite self-energies and vacuum polarisations, local gauge invariance, forcing renormalisation in gauge theories, spontaneous symmetry breaking, permanently confined quarks, colour, just as examples) and of the picture which emerges of the ‘vacuum’ (aether?), as seething with particle-antiparticle pairs of every description and as responsible for breaking symmetries initially present, one can ask whether or not nature is *seriously* supposed to be like that.

One can ask the question but it cannot be answered unless we were to discover independent evidence of another layer of mechanism below that of quantum field theory. Until that time then Quine’s remark is correct: the difference between Homer’s gods and the virtual particles of modern physics *is* one of degree, not kind. But this is only true of the theoretical constructions of fundamental particle physics, not physical objects in general. We should apply different epistemic standards to the higher sciences than those of the bottom level.

---

<sup>6</sup>Both Glennan (1996) and Bhaskar (1978) propose analyses of causation, and its relationship to mechanism, that are consistent with the specific cases discussed here.

<sup>7</sup>Quoted by Cartwright (1983, p7).

### 3.4 Laws and Exceptions

Mercury does not obey Kepler's Laws to the letter, and Newton could not explain why. But Einstein could. If Newton ensured Kepler's place in the book of Physics then why didn't Einstein write Kepler out again? Einstein's explanation of anomalous Mercury certainly *undermined* Kepler, but it does not seem to have been fatal in the way that Kepler's undermining of Ptolemy had been. Why not? The reason is this.

Naturalisation explains the behaviour of a system in terms of the underlying mechanism. The logic of the explanation is thus: if the system works like *this*, then it will behave like *that*. Therefore if the mechanism changes then the behaviour of the system will change. *But the conditional underlying the explanation will still be valid* — it is just the antecedent no longer applies. For example, Newton showed that the planets obey Kepler's laws *if* his laws of gravity and mechanics were accurate. The problem with Mercury is that as it whips round close to the sun then relativistic effects shrink its inertial frame of reference, and the perihelion of its orbit shifts round. Nonetheless it is still the case that if Newton's laws apply, then so will Kepler's.

If the only support we have for a law is its empirical accuracy (or its predictivity) then any counter-example will count as a 'hit' against the law. On the other hand if that law has been naturalised in an underlying mechanism then counter-examples can be accounted for in terms of changes in the mechanism. Of course counter-examples make laws less useful, but there is a difference between being less useful and being disproved. Kepler is not as empirically accurate as Einstein but, given Newton's naturalisation, then it is not disproved. This is why Kepler has been weakened, not deleted.

Consider another example. The periodic table is the most fundamental regularity in chemistry. Mendeleev and Newlands discovered that the properties of the elements showed a periodicity of seven, which enabled Mendeleev to make some of the most startling scientific predictions ever made. He correctly prophesied the discovery and properties of two new elements — gallium and germanium — to fill the obvious gaps in the table, *and* predicted that the accepted atomic weights of tellurium and gold would be found to be wrong because their current values disrupted the otherwise monotonic order<sup>8</sup>. At this point in history the periodicity of the elements — Newlands' 'celestial octaves' — seemed like one of the most powerful universal laws ever discovered. But then came discoveries that were completely unforeseen, and which disrupted the pristine periodic order. First came the sprawling lanthanide's and actinides (which are omitted from most modern copies of the table in order to make the pattern look neater), then came helium and hydrogen which formed a initial period of length two. However the electronic theory later not only explained Mendeleev's periodicity but also accounted for the exceptions. Mendeleev's status is now similar to Kepler's: their immortality does not just rest on the empirical accuracy of the patterns they discovered, but also on the way that they were subsequently naturalised.

The philosophical problem here is that of the epistemological status of *laws*. The traditional view held by most mathematical physicists (with the notable exception of Feynman) is that laws are written, in mathematical language, in the 'Book of Nature' or 'Mind of God'. The first problem with this view — a problem common to all Platonic and dualist schemes — is an ontological one:

---

<sup>8</sup>So proving Rutherford's remark that the correct theory is unlikely to be the one that fits all the facts, since some of those 'facts' are bound to be proved wrong.

what connects the ideal and the actual; what miracle ensures that the objects in our world obey those ideal laws? The second, more pressing, problem is the epistemological one: how do we know what those laws are? Of course empirical evidence can *suggest* the existence of laws but Popper, following Hume, argued that no amount of confirming instances are enough to *prove* a law, even though counter-examples can disprove them. According to Popper we can *never* have knowledge of the laws of nature, all we have are hypotheses that have not been falsified yet. But what does it take to falsify a hypothesis? Have Kepler's and Mendeleev's hypotheses been falsified? No: counter-examples do not necessarily disprove laws. If we can explain the behaviour of the system through naturalisation then we may be able to account for those counter-examples as being due to changes in the working of the underlying mechanism. Exceptions can *prove* rules.

The traditional view of laws went hand in hand with the bias of predictivity: if the correct description of a behaviour is the one that is most predictive then the most significant, or 'real', regularities will be those that are instances of universal laws. However, most laws in the scientific canon, such as the Boyle's or Kepler's, are not 100% accurate or predictive. The usual strategy for dealing with these cases is a mild form of Platonism, in which these laws are said to only apply in an 'ideal' world. Thus although a law, *L*, may not be universally true, it could still be universally true that *L* holds under 'ideal' conditions. So, for example, the gas laws are never 100% accurate and sometimes fail completely, such as when the gas in a cylinder starts to condense. Naturalisation can easily accommodate these counter-examples, but the more traditional strategy is to argue that these laws only apply to an *ideal* gas, not real ones (see, for example, Kripke (1982)). The problem then is to make sense of the connection between events in ideal worlds and events in ours. Fodor, for example, holds that it is only universal laws that are real, not the ideal worlds to which they apply:

*ontologically* I'm inclined to believe that it's bedrock that the world contains properties and their nomic relations; i.e., that truths about nomic relations among properties are deeper than — and hence are not to be analysed in terms of — counterfactual truths about individuals. In any event, *epistemologically* speaking, I'm quite certain that it's possible to know that there is a nomic relation among properties but not have much idea which counterfactuals are true in virtue of the fact that the relation holds. It is therefore, *methodologically* speaking, probably a bad idea to require of philosophical analyses that are articulated in terms of nomic relations among properties that they be, as one says in the trade, "cashed" by analyses that are articulated in terms of counterfactuals among individuals. . . .

Apparently Kripke assumes that we can't have reason to accept that a generalisation defined for idealised conditions is lawful unless we can specify the counterfactuals which would be true if the idealised conditions were to obtain. It is, however, hard to see why one should take this methodology seriously. For example: God only knows what would happen if molecules and containers actually met the conditions specified by the ideal gas laws (molecules are perfectly elastic; containers are infinitely impermeable; etc.); for all *I know*, if any of these things were true, the world would come to an end. After all, the satisfaction of these conditions is, presumably, *physically impossible* and who knows what would happen in physically impossible worlds?

But it's not required, in order that the ideal gas laws should be in scientific good repute, that we should know anything like all of what would happen if there really were ideal gases. All that's required is that we know (e.g.) that if there were ideal gases, then, *ceteris paribus*, their volume would vary inversely with the pressure upon

them. And *that* counterfactual *the theory itself tells us is true*. (Fodor, 1990, p93)

Fodor's criticism of Kripke, that we simply do not know what would happen in ideal worlds, is correct. Many such ideal worlds are indeed physically impossible: if molecules collided elastically then solids, including impermeable gas containers, could not form. It is like imagining a world that contains an unstoppable object and an unmoveable obstacle. Thus we cannot even make proper sense of Kripke's modal counterfactuals, let alone use them as the epistemological foundation of lawhood. However Fodor's alternative to Kripke is circular. Fodor asks what we need to know in order for the gas laws to be "in scientific good repute". His criterion is that the gas laws should hold for ideal gases *ceteris paribus*, and his only justification for believing this is that the theory is true; but a justification for believing this is what we were looking for in the first place. Fodor is 'quite certain that it's possible to know that there is a nomic relation among properties but not have much idea which counterfactuals are true in virtue of the fact that the relation holds', but gives no reason for his certainty.

But we can avoid the metaphysical baggage of modal counterfactuals and ideal worlds if we understand the gas laws as naturalised empirical regularities, rather than universal laws. This only requires that the extent to which molecules and containers approximate the ideal explains the extent to which the gas laws apply. After all, saying that  $x$  tends to  $y$  in the limit does not require that we postulate an 'ideal' point at which  $x$  and  $y$  actually meet. Naturalisation accounts for the observed regularity, and also the exceptions, in a concrete, empirical and metaphysically non-problematic way.

I am *not* inclined to believe that 'it's bedrock' that the world contains properties and their nomic relations; indeed it is hard to make sense of the claim that the world *contains* a law except in a Platonic sense. The world comprises matter whose behaviour exhibits certain regularities, and for this to be true we do *not* need to presuppose prior laws that that matter 'follows' according to its essential nature in some miracle of cosmic obedience. Why does the world of fundamental physics behave as it does? The misleading answer is that it is due to Platonic ideal laws. The honest answer is that we do not know — but the bias of naturalisation warns against turning this necessity into a virtue. This limitation is a peculiarity of the bottom level of physical explanation, and not something that physics envy should tempt us to accept in the higher sciences.

### 3.5 Prediction and Induction

The predictive power of Kepler's theory was not enough, on its own, to save his place in the book of physics. Nonetheless prediction is an essential part of our 'idea of the good' in science and seems to be in some way linked to our ability to explain a phenomenon. The best theories are both naturalisable *and* predictive. So what is the relationship between the two?

The first thing to notice is that naturalisation is not *equivalent* to prediction. The laws of Copernicus or the Mayans are (potentially) as predictive as Kepler's, but only the latter can be naturalised. On the other hand naturalisation does not always yield accurate predictions, for two possible reasons. The first possibility is that although the workings of the system can be understood, they are too complex and sensitive for us to derive predictions in practice. Meteorologists, for example, cannot produce accurate long-term weather forecasts even though there is nothing

mysterious about the mechanisms that drive changes in the atmosphere. The second possibility is that the mechanism underlying the behaviour of the system will itself change. For example, we cannot predict happens to a gas when it condenses just from knowledge of the mechanisms underlying the gas laws.

Naturalisation is not equivalent to prediction. But the bias of naturalisation *does* affect how we use past experience to make predictions. Our experience can be interpreted in many different ways, and different interpretations may generate different predictions. This ambiguity lies behind both Hempel's and Goodman's problems of induction. Hempel's problem concerns the asymmetrical nature of justification and confirmation (1965). Suppose that we were seeking evidence for the inductive claim that "all ravens are black" [ $\forall x(Rx \rightarrow Bx)$ ], of which a black raven [ $Ra \& Ba$ ] is an instance. This claim is logically equivalent to "all non-black things are non-ravens" [ $\forall x(\neg Bx \rightarrow \neg Rx)$ ], which seems to imply, counter-intuitively, that our original claim would be supported by finding a non-black non-raven [ $\neg Ba \& \neg Ra$ ], such as a blue parrot. But the bias of naturalisation implies that our theories about the colour of birds should not just be based on observed correlations, but also by understanding the mechanism that links colour and membership of a species. We can only explain the observed connection between ravenhood and blackness, for example, by understanding the developmental processes connecting the wild-type genome of *Corvus corax* to feather pigment production. This provides good grounds for believing that all organisms that carry those genes would be black<sup>9</sup>. Conversely, explaining the connection between non-black things and non-ravens requires demonstrating a mechanism between being any colour *except* black, and *not* being a living organism carrying that genome. But the only way to do this would be as a logical consequence of having demonstrated the previous connection between ravens and blackness, and blue parrots would be irrelevant for this task.

We can use the same strategy with Goodman's problem of the projectibility of predicates (1955). If we define the property *grue* as being green before the year 2000 and blue thereafter, then we have precisely as much evidence for emeralds being *grue* as green: every instance of an emerald being green in this millennium will also be an instance of one being *grue*<sup>10</sup>. But this implies that we should predict that all emeralds will turn blue at midnight on the 31st December 1999. The reason why we describe emeralds as having a certain constant colour is because we have some intuitions about the mechanisms underlying our observations; in this case it is that the colour of emeralds is due to the way that their crystal structure transmits light. Therefore as long as the mechanism does not change over the millennium then neither will the colour of the crystal. Compare this confidence with our attitude to the millennium computer bug. We may be used to our personal computers working happily, but because we have some knowledge about how they store and process dates, and how this mechanism will be affected by the increment from '99' to '00', then we intuit that, unlike emeralds, their behaviour may well change when the millennium comes.

Naturalisation provides a guide as to which predictions we should draw from our observations, but it also gives us clues about which predictions we should not. If we know that an observed regularity is coincidental, and *not* due to a similarity in the underlying mechanism, then we are less likely to lay bets on it persisting. Chairs, for example, come in many different materials,

<sup>9</sup>This example will be significant when discussing the heritability of biological traits in chapter 8.

<sup>10</sup>This simplified version of the original problem is due to Gärdenfors (1990).

shapes, and sizes. They need have nothing in common other than their ability to provide a seat. It may be the case that *every* chair we have sat on weighed about 10lb, but this may have been for a different reason in each case: one chair may have been made of wood, another of metal, and so on. Therefore we know that there is nothing in the nature of chairs to make them always weigh 10lb. It is likely that other chairs we come across will be similar to the ones we have seen before, but we would *not* be particularly surprised to come across an inflatable armchair that weighed only a few ounces or a throne that weighed a ton — it would not cause us to rethink what chairs are. On the other hand we would be puzzled to come across a chair that was not suitable for sitting on. Would it *really* be a chair? But the reason why we predict that all chairs can be sat on is due to how we define what a chair is, and not from inductions about chairs that we have encountered.

Naturalisation also provides a way of accounting for predictions that do not succeed, just as it could provide a way of accounting for exceptions to laws. The Victorians, for example, were surprised to discover swans in Australia that were black, rather than white — just as we would be surprised to discover an albino raven. But does this mean that the Victorians were foolish to predict that all swans were white? No, because the black swans belonged to a new sub-species which, like the albino raven, had a slightly different genetic make-up to those previously observed. Therefore the developmental mechanism on which they based their predictions had changed. The prediction was just as valid as before, it is just the scope of its application that had to be revised.

I agree with Goodman that ‘the problem is not to guarantee that induction will succeed in the future — we have no such guarantee — but to characterise what induction *is* in a way that is neither too permissive nor too vague’<sup>11</sup>. Naturalisation does not guarantee that a prediction will succeed, but it does explain how we may produce predictions on the basis of past experience. The important point is that confident predictions are not just based on the accumulation of empirical evidence but on knowledge, or intuitions, about the mechanism underlying that evidence.

### 3.6 Conclusion

Naturalisation embodies a certain ‘idea of the good’ in science. It is a way of sorting through all the possible descriptions of our empirical evidence in a way that (1) explains why the world behaves like that, and (2) also explains why sometimes it does not. It is an idea of the good that Kepler and Darwin and Mendeleev and Mendel lived up to, but Ptolemy did not.

Now when we talk about the ‘great’ theories of science we usually think of the revolutions in fundamental physics, of Newton and Einstein and Quantum Mechanics. But because they were concerned with the bottom level of nature then naturalisation is not an option for these theories, and so different criteria of goodness apply. Unfortunately physics envy has meant that the latter ideal is held up as the standard that the rest of science should aspire to. This is a mistake, and we shall see some of the implications of this mistake in the rest of this thesis.

---

<sup>11</sup>From Putnam’s foreword to the fourth edition of *Fact, Fiction and Forecast*.