

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 geocentric astronomy based on epicycles. If enough circles were stacked upon each other then the geocentric 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 three¹ 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

¹The 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 α -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 future². 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

²The 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.³

(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.

³This 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 system⁴. 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

⁴Actually 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 tautologous⁵ 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-

⁵This 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 here⁶, 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)⁷,

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.

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

⁷Quoted 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 order⁸. 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:

⁸So 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 black⁹. 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*¹⁰. 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,

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

¹⁰This 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’¹¹. 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.

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