The Mind–Body Problem and Explanatory Dualism

NICHOLAS MAXWELL

1 Introduction

One important part of the mind–body problem arises because it seems impossible that a scientific account of what goes on in a conscious brain, however complete, could of itself predict the conscious experiences of the person whose brain it is. Ordinarily, perhaps, we are not too puzzled by the fact that we have inner experiences. Invoke science, to arrive at a better explanation and understanding of inner experiences, and we encounter neurons, synaptic junctions, exchange of potassium and sodium ions across semi-permeable membranes and so on, but never anything, apparently, remotely like a sensation, a feeling, a conscious experience. The better the scientific explanation, the more inexplicable our inner experiences seem to become, the more they seem to disappear.

It is this apparent inherent resistance of mind to scientific explanation, the apparent stubborn scientific unintelligibility of mind, that engenders an important part of the mind–body problem.

Traditional dualism just postulates that there is this mysterious entity, the mind, that is, mysteriously, beyond the reach of science. Behaviourism, the identity theory, and various versions of functionalism, postulate that, despite appearances to the contrary, nothing mental exists that *is* in principle beyond the scope of scientific explanation. In this paper I argue for a version of the two-aspect theory: perceptual qualities of things external to us, and mental aspects of brain processes, really do exist and are beyond the scope of science. However, consideration of what scientific explanation can be expected to achieve, even at its most optimistic, reveals that it is entirely unreasonable to expect that even a full scientific explanation of everything could explain the sorts of things that we may suppose mental phenomena to be. Even a complete physical explanation of everything, in terms of the yet-to-be-discovered true physical theory of everything, would be designed to refer to, describe and explain only a highly selected aspect of all that there is (or might be). Precisely because even a complete physical account of the world would pick out only one very special *kind* of feature of things, it is unreasonable to expect that such an account would tell us everything about everything. In short, the inherent resistance of

https://doi.org/10.1017/S003181910000005X Published online by Cambridge University Press

the mental to physical explanation is due, not to some built-in unintelligibility of the mental, but to built-in limitations of physical explanation. There is, in other words, an *explanation* as to why the mental cannot be scientifically explained.

But if the mental cannot be understood scientifically, how is it to be understood? There is another kind of explanation, I shall argue, which may be called 'personalistic' explanation. This is an entirely respectable kind of explanation; it works, however, in a certain sense, in the opposite direction to scientific explanation. As things become increasingly *personalistically* intelligible, they become, roughly, increasingly *scientifically* unintelligible, and vice versa. As the contents of a conscious person's head come increasingly into focus *scientifically*, as a brain or physical system, inevitably the mental aspects seem to disappear; as the contents of the person's head come increasingly into focus *personalistically*, as a mind, so the brain, the neurons, the physical system seem to disappear. The key to solving this important part of the philosophical or conceptual mind–body problem is to recognize a dualism, not of kinds of entity, but of kinds of explanation.

2 Theoretical Physics

My claim is that the proper ultimate task of theoretical physics, at its most ambitious, is to predict and explain, not everything about everything, but at most only a highly selected *aspect* of what there is. The task is to discover the true theory of everything, T, which (a) unifies all forces, fields and particles,[1] (b) applies in principle to all phenomena, and (c) in principle predicts and explains all phenomena in the sense that, given any isolated system (possibly an instantaneous state of the entire universe), T, together with a precise specification of the state of the system at some instant t couched in the vocabulary of T, suffices (in principle) to imply specifications of all subsequent (and prior) states of the system when described with the same vocabulary, there being no loss of content in these predictions, the presumption being that the system remains isolated, and that the universe is deterministic.[2]

[1] That it is reasonable to hold that the universe is such that there is underlying physical unity, the universe being physically comprehensible, is argued in my *The Comprehensibility of the Universe: A New Conception of Science* (Oxford University Press, 1998).

[2] All sorts of qualifications and modifications need to be added to this to take into account such things as probabilism, the non-existence of isolated systems, field theories and other theories of modern physics such as special

https://doi.org/10.1017/S003181910000005X Published online by Cambridge University Press

In order to be complete in this sense, T must satisfy *two* conditions. First, it must apply to everything, to all possible isolated systems. The vocabulary of T must be sufficiently rich to specify the precise instantaneous state of any isolated system, or the instantaneous state of the universe. Second, T must specify precisely all the forces that there are, all the kinds of interaction, so that the specified predictive task can in principle be performed.

Physical features, in this sense, are features which (as far as possible) everything has in common with everything else, and which are *causally efficacious* in the sense that they determine the way things change. In order to be complete, T must specify precisely all such actual physical features.

The decisive point to appreciate is that completeness in this sense does not mean completeness in the sense that T would predict *everything* about everything, everything that is true about all isolated systems. If an isolated system has features which do not need to be described in order for the predictive task indicated above to go through, then T will make no mention of such features. If these features are such, furthermore, that descriptions of them are not entailed by any descriptions couched in the vocabulary of T, then these features will be *non-physical*, lying outside the scope of even a complete theoretical physics. If there are non-causally efficacious features that have to do with what things look like or feel like, with what it is like to *be* something, or with what things *mean*, and the world is such that T does not need to refer to or describe these features in order to fulfil the above predictive task, then it won't. The basic task of theoretical physics is such that it remains silent about such features, even though they exist. Thus, the fact that physics *is* silent about such features—colours, sounds and smells as we perceive them, inner experiences, the content of our thoughts and utterances—provides no grounds whatsoever for holding that such features don't exist, or are inherently unintelligible if they *do* exist. A complete *physical* description was never intended to be a *complete* description.

It is only if causally efficacious features of things are the only kind of feature that there is, that it would be the case that the *physical*

---

and general relativity and quantum theory, and practical restrictions on prediction stemming from such things as the impossibility of specifying the precise physical state of even the simplest system, and the impossibility of solving the equations of physical theories except for the simplest systems, and then often only approximately. None of these qualifications have a bearing on the argument of this paper and so are here ignored. (We are, for example, concerned with what T can, ideally, predict or imply, not with what we humans can actually do with T.)

https://doi.org/10.1017/S003181910000005X Published online by Cambridge University Press

completeness of T would render T wholly complete and comprehensive. But why should causally efficacious properties be the only kind of property to exist? Experience and common sense indicate, in my view correctly, that the world is much fuller and richer than a world denuded of everything but the causally efficacious.

3 Science

The point just made—that the physically complete need not be complete—may seem to some to be a triviality. Even in physics, it may be argued, there are laws and theories, such as those of statistical mechanics and thermodynamics, which correctly apply to phenomena, but which contain concepts (such as probability, temperature or entropy) not contained in current fundamental physical theory, and hence unlikely to be contained in the true theory of everything, T. And if we add on chemistry, and other parts of natural science, the point becomes even more blatant. Far from embracing everything, physics does not even include all of natural science.

This argument is not valid. It is easy to see how it is possible for there to be a law or theory L which (a) contains concepts that are not a part of the theory of everything, T, and yet (b) does not assert anything true that is not derivable from T. As an elementary toy model for this, let T be 'All objects are spheres' and L be 'All objects are ellipsoids' (spheres being a special kind of ellipsoid). If T is true, then so is L; it is not possible for T to be true and L to be false.[3] Thus, even though L contains a concept not included in T, what L asserts truly can be derived from T. Phenomenological and macroscopic laws and theories of physics, in so far as they are true, are similar: they employ concepts not included in T, but make

[3] In this toy case, and in all such more realistic cases, bridge statements that take one from theoretically described to phenomenologically described entities are analytic, whereas bridge statements that take one in the reverse direction are empirical. This is a consequence of phenomenological terms and laws being merely a less contentful way of describing theoretical entities and processes. Thus 'If this object is a sphere it is an ellipsoid' is analytic, whereas 'If this object is an ellipsoid then it is a sphere' is at most a contingent truth. Just this case that needs to be considered in realistic examples of the reduction of one theory or science to another is excluded *a priori* by some contributors to the philosophy of mind, in that bridge statements are held to be *biconditionals*, that are either analytic, or contingent: see, for example, J. Kim, *Philosophy of Mind* (Boulder: Westview Press, 1996), 213–6.

https://doi.org/10.1017/S003181910000005X Published online by Cambridge University Press

assertions sufficiently (a) restricted in scope, and (b) imprecise, to be both true and derivable from T.[4]

In short, in so far as natural science is concerned with the causally efficacious, or that which can be reduced to the causally efficacious, the mere fact that there are natural sciences that employ concepts not found in fundamental theoretical physics does not provide grounds for holding that the true physical theory of everything would be scientifically incomplete.

4 The Experiential

Let us now consider an isolated system that is a candidate for containing things and processes that have non-physical features. It consists of a space capsule which, in turn, contains a conscious, experiencing person. Physical descriptions of instantaneous states of the system at times t (couched in the vocabulary of T), will of course include complete specifications of the physical states of the person's brain, body and environment. But this does *not* mean that these (T-based) physical descriptions will cover all features of things in the isolated system. The colours, sounds, smells, tastes, tactile qualities that the person experiences; the inner sensations, feelings, thoughts, desires and imaginings of the person; and what the person says or writes or does: these experiential and personalistic features will not need to be included in T-based descriptions as long as the above predictive task is not thereby impaired. T-based descriptions will of course describe physical processes associated with such experiential features, such as light of diverse wavelengths being absorbed and reflected by such and such physical objects, potassium and sodium ions passing through physical structures that are the surface membranes of neurons of the person's brain (associated with perceptions, feelings and thoughts); physical processes associated with vibrating vocal chords, or with limbs being moved. Physical completeness does not require, however, that the *experiential* or *personalistic* features of all this be mentioned.

[4] This account of a law, L, asserting nothing in addition to, and being reducible to, a more fundamental theory, T, runs into the difficulty that often in the kind of situation we are envisaging, L is incompatible with T. (Thus Kepler's laws of planetary motion, and Galileo's laws of terrestrial motion are incompatible with Newtonian theory, which is in turn incompatible with Einstein's general theory of relativity.) This difficulty can be overcome by employing the solution to the problem of verisimilitude, and the solution to the problem of scientific realism, expounded in chapter six of op. cit. note 1.

https://doi.org/10.1017/S003181910000005X Published online by Cambridge University Press

It is of decisive importance to note that, in so far as I want to know about the experiential and personalistic aspects of what is going on in the capsule, I want (and need) quite essentially to relate the sentient being in the capsule to *myself*. I need to bring myself into the picture, in a way in which I must not do if I seek a physical description and explanation of what is going on inside the capsule. Suppose the conscious being inside the capsule is an alien. If I want to know what the interior of the capsule looks like to the alien, I want to know what it would be like for me to have occur in my brain processes that are similar in relevant respects to the processes that are going on inside the alien's brain, associated with perception. And similarly, if I want to know what the alien is experiencing or feeling, I want to know how it would be for me if processes similar in the relevant respects to the processes going on in the alien's brain were to occur in my own brain. (It may not be possible, of course, for me to know any of this because my brain is too different from the alien's brain.) If I want to know what the alien asserts, writes or thinks, I want to know what these assertions or thoughts are when translated into my language. All experiential or personalistic aspects of things in the capsule bring *me* into the picture in an essential way, and involve knowing how things in the capsule relate to *my* experiences and thoughts. Physical descriptions, explanations and understanding of what is going on inside the capsule, however complete, at no point involve or provide this kind of anthropomorphic, *personalistic* information: it is this which ensures that the personalistic, the experiential, cannot be reduced to the physical.

An elementary argument establishes that *physical completeness* cannot be *completeness*—or, in other words, that purely *experiential* features of things cannot be *physical* features. All physical properties are such that it is not necessary to have any special kind of experience in order to know what sort of properties they are. In order to know what 'mass', 'charge', 'energy' or 'spin' *mean* it is not necessary to have any special sort of experience. In particular, being blind from birth, so that one has never had any visual experiences, does not debar one from understanding the physical theories of optics—classical or quantum mechanical—just as well as anyone else. But when it comes to experiential properties, such as colours as we experience them, then it *is* necessary to have had special sorts of experiences oneself in order to know what sort of properties these are. In order to know what sort of thing redness (or the visual experience of redness) is, it *is* necessary, at some time in one's life, to have had the visual experience of redness. A person completely colour-blind from birth cannot know what sort of thing redness (or the experience of redness) is.

https://doi.org/10.1017/S003181910000005X Published online by Cambridge University Press

This argument is usually attributed to Thomas Nagel and Frank Jackson.[5] It was in fact spelled out by me in a paper published in *The British Journal for the Philosophy of Science* in 1966, and in a paper published in *Australasian Journal of Philosophy* in 1968, the first of which appeared eight years before Nagel's paper, and sixteen years before the first of Jackson's papers.[6]

One unfortunate consequence of the neglect of these papers of mine, (quite apart from adverse consequences for my subsequent academic career!) is that one crucial point that I sought to communicate in them has, I feel, over thirty years later, still not been adequately grasped. It is one of the basic points of the present paper (to be developed further below), and can be put like this. The fact that science does not and, it seems, *cannot* predict sensory and experiential features of things external to us and brain processes within us is no grounds whatsoever for holding that such sensory or mental features are inherently unintelligible if they exist, or nonexistent if one holds that the inherently unintelligible does not exist. Philosophers with as divergent views as Nagel or McGinn on the one hand, and Dennett on the other,[7] unite in overlooking this simple point. In effect they agree that the irredeemably experiential (if it exists) is inherently scientifically unintelligible. Nagel and McGinn do not think this constitutes adequate grounds for concluding that the experiential does not exist, while Dennett does. What both parties overlook is that science is not intended or designed to predict the experiential; its failure to do so does not at all imply that the experiential is inherently inexplicable.[8]

I must emphasize that the above argument, as I first spelled it out in '66 and '68, does not just seek to establish that the silence of

[5] I refer, of course, to T. Nagel, 'What Is It Like to Be a Bat?', *The Philosophical Review* **83**, No. 4 (1974), 435–50; F. Jackson, 'Epiphenomenal Qualia', *Philosophical Quarterly* **32** (1982), 127–36; F. Jackson, 'What Mary didn't Know', *Journal of Philosophy* **83** (1986), 291–5.

[6] See N. Maxwell, 'Physics and Common Sense', *British Journal for the Philosophy of Science* **16** (1966), 295–311—see especially 303–8; N. Maxwell, 'Understanding Sensations', *Australasian Journal of Philosophy* **46** (1968), 127–45—see especially 127, 134–7 and 140–1. When I recently drew Thomas Nagel's attention to these publications, he remarked in a letter, with great generosity: 'There is no justice. No, I was unaware of your papers, which made the central point before anyone else'. Frank Jackson acknowledged, however, that he had read my 1968 paper.

[7] See T. Nagel, *The View From Nowhere* (Oxford University Press, 1986); C. McGinn, *The Problem of Consciousness* (Oxford: Blackwell, 1990); D. Dennett, *Consciousness Explained* (Boston: Brown, Little and Co., 1991).

[8] In addition to the two papers already mentioned, see N. Maxwell, *From Knowledge to Wisdom* (Oxford: Blackwell, 1984), 259–64.

https://doi.org/10.1017/S003181910000005X Published online by Cambridge University Press

physics about the mental aspect of brain processes gives us no grounds whatsoever for supposing that this aspect does not exist; just as emphatically, it seeks to establish that the silence of physics about perceptual properties of things around us as experienced by us, the greenness of grass, the redness of roses, provides no grounds whatsoever for holding that these features don't really exist out there in the world.

But surely, it may be objected, colours as we experience them are only *subjective*; they are not *objective* features of things out there in the world! My answer to this has not changed much since '66. There are two quite different ways of drawing the distinction between the subjective and the objective. In terms of one distinction, colours are objective; in terms of the other distinction, colours are subjective.

We may say, on the one hand, that a property P is 'existentially objective' if it exists and 'existentially subjective' if it only appears to exist but in reality does not. And on the other hand, we may say that P is 'humanly objective' ('humanly subjective') if it is not necessary (*is* necessary) to have a special sort of experience to discover what sort of property it is.[9] Colours, I claim, are existentially objective and humanly subjective. They really do exist out there in the world, but in order to discover what sort of features of things they are, you need to have the right kind of sense organs and nervous system to be able to perceive them. Conscious beings from other planets (and to some extent other sentient beings from this planet) are no doubt aware of all sorts of perceptual qualities of things that we know nothing of. (The mere *possibility* of there existing sentient, conscious beings with sense organs and physiologies different from ours suffices to ensure that things have perceptual qualities of which we can know nothing.)

The above argument for the incompleteness of physics (and physicalism) has, of course, been criticized and rejected. Dennett, for example, argues that Jackson's Mary, prevented from ever seeing colour, might nevertheless know, when presented for the first time with a blue banana, that the colour is wrong.[10] But, in explaining how she knows this, Dennett is forced to acknowledge that she has experimented with the effect of light of various wavelengths on her own nervous system: either, in the past, she has induced in herself the relevant visual sensations (which is cheating); or she has investigated the brain processes of others experiencing colour, and has activated a device which tells her what sort of brain process is occurring in her own brain as she looks at the blue banana. If the latter, then she

[9] For this second way of distinguishing objective and subjective see my 'Physics and Common Sense', 310–1.

[10] See D. Dennett, *Consciousness Explained*, pp. 398–401.

https://doi.org/10.1017/S003181910000005X Published online by Cambridge University Press

could, more straightforwardly, discover that the banana is the wrong colour by checking, by means of a physical instrument, what range of wavelengths of light the banana reflects. She can remain completely colour blind, and still know that the banana is the wrong colour. But in this case, of course, she would remain ignorant of what sort of thing blueness, or yellowness, as perceived, *is*. Dennett's counter-argument does not begin to come to grips with the above argument for the incompleteness of physicalism.

In what follows, in any case, I assume that the Maxwell-Nagel argument (as I may perhaps call it[11]) is valid. If there really are features of things which are such that, in order to know what sort of features they are, it is necessary oneself to have a certain sort of experience, and hence (we may presume) a certain sort of neurological process occur in one's own brain; and if, in addition, no mention of these features needs to be made in order to carry out the kind of predictive task described in section 2 above, then such features will lie irredeemably beyond the scope of physics.

But why cannot physics be extended to include reference to experiential features of things? This will be considered and rejected in section 8 below.

5 Personalistic Explanation

The fact that experiential or mental features of things or brain processes are beyond the scope of scientific explanation does not mean that these features are inherently inexplicable, for these features can be explained and understood personalistically—or must be presupposed to be intelligible by personalistic explanation in that they are included in the explanans of personalistic explanations. The visual sensation of redness, utterly inexplicable scientifically, is wholly understandable personalistically (at least for normal sighted persons.)

Personalistic explanation is a kind of explanation that is entirely valid, intellectually fundamental (when viewed from a certain perspective) and irreducible to scientific explanation.

Personalistic explanations seek to depict the phenomenon to be explained as *something that one might oneself have experienced, done, thought, felt*. In seeking personalistic understanding of another person, or being, I want to know how it would be for me to *be* the other person, thinking, feeling, hoping and fearing, seeing, imagining and doing what the other person thinks, feels, etc. As I have already

[11] Jackson informs me that he now no longer believes the argument to be valid (personal communication).

https://doi.org/10.1017/S003181910000005X Published online by Cambridge University Press

indicated, there is an irredeemable *anthropomorphic*, even *personal* dimension to personalistic understanding: it always involves understanding the other by using oneself as a model of the other, and rearranging, in one's imagination, aspects of oneself, such as one's circumstances and environment, character, beliefs, experiences, values, goals so that these come to be the circumstances etc., of the other person. One understands the other person by becoming the other person in one's imagination and then understanding the new self one has imaginatively become. Or, in other words, personalistic understanding involves *imitating* the other in imagination, the other's inner doings (thoughts, experiences, feelings and so on) being imitated in imagination as well. One then seeks to understand the other by understanding the self that has been constructed by imaginatively imitating the other, just as one would ordinarily understand one's self.[12] It is sometimes called *empathic* understanding or, by psychologists (usually with dismissive connotations), understanding of *folk psychology*.[13] Personalistic understanding is the kind of understanding we (more or less) have of ourselves and of others in our life; it is the stuff of biography, history, anthropology, literature, psychotherapy.[14]

6 The Intellectual Authenticity of Personalistic Explanation

The claim that the mental aspect of brain processes, though incomprehensible scientifically, is nevertheless genuinely comprehensible personalistically, is only valid if personalistic explanation is a fully authentic mode of explanation in its own right, one that really does render mental phenomena comprehensible.

Within academia, however, there is a tendency to regard personalistic understanding as inherently intellectually crude and primitive: psychologists and philosophers who call personalistic understanding

[12] For a more detailed account of the conception of personalistic explanation that I am employing here, including a sketch of its evolution, see op. cit. note 8, 174–81, 183–9, 264–75.

[13] S. P. Stich, *From Folk Psychology to Cognitive Science: the Case Against Belief* (Cambridge, Mass: MIT Press, 1983); P. M. Churchland, Eliminative Materialism and Propositional Attitudes, *Journal of Philosophy* **78**, 1981, pp. 67–90.

[14] Personalistic explanation, as understood here and as characterized more fully in op. cit. note 8, differs in important respects both from the theory theory of folk psychology and from the simulation theory, for which see: P. Carruthers and P. Smith (eds) *Theories of theories of mind* (Cambridge University Press, 1996); M. Davies and T. Stone (eds) *Folk Psychology: The Theory of Mind Debate* (Blackwell, 1995).

https://doi.org/10.1017/S003181910000005X Published online by Cambridge University Press

'folk psychology' tend to hold such an attitude.[15] (Folk psychology is thought of as being rather like pre-Galilean physics, highly primitive and in urgent need of being replaced by something more adequate, as pre-Galilean physics has been replaced by modern academic physics.)[16] What is at issue is not how good or poor the intellectual quality of this or that attempt at personalistic understanding is, but rather whether *all* personalistic understanding is inherently poor intellectually.

One reason why the intellectual quality of personalistic understanding may be so poorly thought of, in this way, is that this type of understanding fails to meet *orthodox* criteria of intellectual excellence, to a quite dire extent.[17] Viewed from the standpoint of

[15] See note 13 for references.

[16] 'The term 'folk psychology' is … intended to portray a parallel with what might be called 'folk physics', 'folk chemistry', 'folk biology', and so forth.' P. M. Churchland, 'folk psychology (2)', in S. Guttenplan (ed.) *A Companion to the Philosophy of Mind* (Oxford: Blackwell, 1994), p. 308. For additional literature on the nature and status of folk psychology, see: J. Greenwood (ed.) *The Future of Folk Psychology* (Cambridge University Press, 1991); B. von Eckardt, 'folk psychology (1)', in S. Guttenplan, op. cit., pp. 300–7.

[17] Objections to the genuineness of personalistic explanations that I consider here are rather different from Churchland's criticisms of folk psychology. In part this is due to the fact that personalistic explanation, and folk psychology as construed by Churchland, are different conceptions. Even so, it needs to be shown that personalistic explanation does not succumb to Churchland's arguments against folk psychology. Churchland argues for the falsity of folk psychology on the grounds that (1) it fails to explain a variety of central psychological phenomena, such as mental illness, sleep, creativity, memory, intelligence differences, and the many forms of learning; (2) it has failed to be empirically progressive since the ancient Greeks 2500 years ago; and (3) it fails to be integrable with the rest of natural science: see P. M. Churchland, op. cit., pp. 310–1. My reply to this, interpreted as a criticism of personalistic explanation is, briefly, as follows. (1) The point of personalistic explanation is to enable us to understand others and ourselves *as persons*, not as physical, neurological or biological systems. Personalistic explanation needs, of course, to be supplemented with scientific explanation. For a magnificent example of the way in which the two kinds of explanation can, and need to, work in tandem see O. Sacks, *Awakenings* (Harmondsworth: Penguin, 1976). (2) I would argue that there has been a great increase in the range, depth, sensitivity and accuracy of personalistic understanding since the ancient Greeks, among those best at the art of such understanding, especially when pursued by someone like Sacks, and especially in connection with modern awareness of the role of unconscious motivation and emotion. But to suppose that personalistic explanation will develop like an empirical science is to

https://doi.org/10.1017/S003181910000005X Published online by Cambridge University Press

orthodox, standard empiricist conceptions of what it is to be scientific,[18] personalistic understanding must be judged to be intellectually crude and primitive in the extreme when compared to physical understanding. Physical understanding is (a) objective (b) impersonal (c) factual (d) rational (e) predictive (f) testable and (g) scientific, in that there is an objective, impersonal, factual theory, which predicts the phenomenon to be understood, and is independently testable, and so amenable to being appraised scientifically and rationally. Personalistic understanding, by contrast, may be held to be (a) subjective (b) personal (c) emotional and evaluative (and thus non-factual) (d) intuitive (and thus non-rational) (e) non-predictive, and (f) untestable. Judged in terms of orthodox scientific standards, personalistic understanding is an intellectual disaster.

Elsewhere I have argued at length that the fault lies, not with personalistic understanding, but with orthodox intellectual standards.[19] Not only do we need a new conception of science, which sees science as being obliged to make problematic assumptions about the nature of the universe, the aim and methods of science evolving with evolving knowledge.[20] More generally, we need to adopt and put into practice a new conception of rational inquiry, according to which the proper fundamental intellectual aim of inquiry as a whole is to help promote *wisdom* by co-operatively rational means (wisdom being the capacity to realize what is of value in life for oneself and others). Inquiry of this type, rationally designed to promote wisdom, would give intellectual priority to (i) articulating our problems of living, and (ii) proposing and critically assessing possible co-operative actions from the standpoint of their capacity to lead to the realization of what is of value. But in doing (i) and (ii) as far as a particular person is concerned, we *are* acquiring personalistic understanding of that person. Giving intellectual priority to tackling problems of living in co-operatively rational ways *is* giving intellectual priority to

[18] For expositions of standard empiricism see op. cit., note 8, 21–3; and op. cit., note 1, 2–3 and 37–45.

[19] See op. cit. note 8, 181–9 and the rest of the book, for the argument that personalistic understanding is intellectually fundamental when viewed from the standpoint of the philosophy of wisdom, but apparently intellectually disreputable when viewed from the defective views of standard empiricism and the philosophy of knowledge.

[20] For this part of the argument see op. cit., note 1. For a summary of the argument see N. Maxwell, 'Has Science Established that the Universe is Comprehensible?', *Cogito* **13** (1999), 139–45.

misconstrue its nature and use. (3) Personalistic explanation is compatible with, but not reducible to scientific explanation; its evolution, however, deserves to be studied within biology, evolutionary theory, ethology, anthropology and history.

https://doi.org/10.1017/S003181910000005X Published online by Cambridge University Press

the development of personalistic understanding. According to the philosophy of wisdom (which depicts a kind of inquiry rationally devoted to the pursuit of wisdom) personalistic understanding is intellectually fundamental; it is essential for co-operative rationality; it is involved in all other branches of inquiry, even the most impersonal such as logic or theoretical physics; the whole of academic psychology and philosophy rests upon personalistic understanding. There can thus be no question of the intellectual standing or adequacy of this mode of understanding. Personalistic understanding at its best is: (a) objective (b) inter-personal (c) emotional and evaluative but also factual (d) intuitive but also rational (e) predictive in a loose way (f) capable of being assessed rationally (e.g. critically) and in terms of human experience.

It deserves to be noted that there is a sense in which physical and personalistic explanations work in opposite directions. Whereas personalistic explanation explains by reducing the unfamiliar to the familiar, physical explanation does almost the opposite: it explains by reducing the familiar to the unfamiliar. Personalistic explanation explains by revealing unfamiliar experiences and actions to be elaborations of familiar, intelligible, rational experiences and actions. The more intelligible something becomes personalistically, the more fiendishly complex and difficult to understand it becomes physically—in that it involves brain processes, even infinitely many different possible brain processes. The more intelligible something becomes physically, the more incomprehensible it becomes personalistically, the physically intelligible being some simple, elementary system exemplifying the unified pattern of the true physical theory of everything in a simple fashion, and thus being remote from personalistically intelligible human experience. There is a sense, then, in which the two kinds of explanation are mutually exclusive, and work in opposite directions: the more intelligible something becomes in one way, the more unintelligible it becomes in the other way, and vice versa. This point contributes to the solution to the mind–brain problem being proposed here in accounting for the way in which understanding head processes scientifically as neurological or physical seems to exclude the very possibility of understanding these processes as mental or conscious, and vice versa.[21]

[21] McGinn and Nagel both hold, as we have seen, that consciousness may ultimately be unintelligible; they come to this (in my view mistaken) conclusion because they fail to appreciate that there are two different kinds of explanation, which work in opposite directions, consciousness being irredeemably incomprehensible in terms of one (scientific) mode of explanation, but comprehensible in terms of the other (personalistic) mode of explanation. Consciousness seems incomprehensible because the mode of understanding which renders it comprehensible is not taken sufficiently seriously.

https://doi.org/10.1017/S003181910000005X Published online by Cambridge University Press

7 Psycho-Functionalism

The experiential (or mentalistic) feature of a neurological process is that feature which one becomes aware of if the process occurs in the right kind of way in one's own brain. If I am to apprehend the experiential features of a neurological process, P, going on in someone else's head, then a sufficiently similar neurological process, Q, must occur in my head, causally and functionally related to the rest of my brain in a way that is sufficiently similar to the way P is related to the rest of the other person's brain. Immediately the question arises: What does 'sufficiently similar' mean here? There are various possibilities.

(i) P and Q are precisely the same physically, even if the two brains are not precisely the same.

(ii) P and Q are precisely the same neurologically (i.e. the same pattern of neurons fire in the same way), even though there are otherwise differences between the physical states of the neurons.

(iii) Neurons may be quite different physically (e.g. in one case neurons are biological, in the other case made out of microchips), but the pattern of firing of the neurons, and the interconnections between the neurons, is the same.

(iv) 'Strength of signal' may be coded in quite different ways at the neuronal level (so that in one case this is related to rapidity of firing of neurons, while in the other case it is related to strength of electric current, let us suppose); once these differences are ignored, however, the pattern of signals is the same in the two cases.

(v) The *functional* or *control* role of the neurological processes, P and Q, are identical in the two brains, even though the pattern of signals, the 'code' at the neuronal level, and the physical structure and functioning of the neurons, are entirely different.

(vi) The behaviour of the two beings is similar, even though the control architecture of the two brains is entirely different so that, from a *functional* or *control* standpoint, the neurological processes, P and Q, work in quite different ways.

As an example of (vi), consider the somewhat fanciful possibility that there is a robot which is controlled by a computer which contains a model of my living brain. The robot's computer brain calculates how my brain would behave in such and such circumstances, one step at a time, and on the basis of the result, gets the robot to act accordingly. The processes going on in the robot's brain are nothing like the processes going on in my brain, even when described in control or functional terms; for one thing, there is a massive amount of parallel processing going on in my brain (which in part accounts for the richness of the content of consciousness); in

https://doi.org/10.1017/S003181910000005X Published online by Cambridge University Press

the robot's brain everything is done sequentially, one step at a time. Thus, even though my behaviour, and that of the robot, are (by hypothesis) the same, nevertheless the processes going on inside our respective brains are entirely different, even when described in purely control or functional terms. (My *brain* does not exist in the robot's computer brain, only a *model* of my brain: and as Searle reminds us, a *model* of a brain is not a brain, any more than a model of a snowstorm is a snowstorm.)

If we adopt (i) we should have to conclude that we cannot understand each other's inner experiences; we should probably have to conclude that we cannot understand our own inner experiences that we have on different occasions. If we adopt (ii), it becomes possible to hold that we human beings have common inner experiences, but no robot can have inner experiences like ours. If we adopt (iii), we can make sense of the idea that a robot has the same kind of inner experiences that we have. Adopting (iv) or (v) ensures that a wider class of robots have experiences like ours, whereas adopting (vi) ensures that any being, however constituted, that behaves *as if* it has inner experiences like ours, thereby *does* have experiences like ours.

How are we to decide between (i) to (vi)? We have an intuitive idea of what we mean when we say that another person's visual sensation of redness is the same as ours: does this correspond to 'sufficiently similar neurological processes going on in our heads' in sense (i), (ii) … or (vi)? (i) is implausible. (vi) is a version of functionalism scarcely distinguishable from behaviourism; it deserves to be rejected for the same reasons as behaviourism deserves to be rejected. This leaves (ii), (iii), (iv) and (v). It may be that understanding more about the neurological nature of our inner experiences will put us into a better position to choose between these four options. My inclination is to plump for (v), while at the same time holding that conscious robot brains are *not* in practice possible, because consciousness requires there to be an incredibly subtle relationship between the *structure* of the brain and the way it *functions* which can only come into existence as a result of a kind of *growth* that is responsive to the way the brain *functions*. If this conjecture is correct, then only those brains are conscious which support growth, and which are therefore, to that extent at least, *biological* in character.

Adoption of (v) amounts to the adoption of a view that may be called *psycho-functionalism*. According to psycho-functionalism, the mental aspect of a brain process is that aspect you become aware of when the process occurs appropriately in your brain, or when a functionally similar brain process occurs in a functionally similar brain, in the sense of (v) above. Mental states and processes map on

https://doi.org/10.1017/S003181910000005X Published online by Cambridge University Press

to appropriate *functionally described* brain states and processes, in the sense of (v).

8 Expanding Physical Explanation

We have seen above that even the true theory of everything, T, would be silent about the mental, the experiential aspect of things. At once it may be asked: why should not physical explanation be expanded, in some way, so as to include the experiential?[22] Why should not additional postulates be added to T to form T*, let us say, where T* predicts the existence of experiential and personalistic features in addition to physical features?

The answer is that such a move would entirely destroy the *explanatory power* of T. In order to turn T into a *complete* theory, T*, postulates will need to be added to T that correlate complex physical states of affairs with all possible experiential features. Each of these postulates will be quite incredibly complicated. In order to specify the physical state of affairs that correlates with redness_e, for example (where redness_e is redness as experienced by us), it is quite insufficient to specify the immense range of molecular structures which absorb and reflect light of wavelengths which lead us, in ordinary circumstances of illumination, to see the objects in question as red_e. In addition, we must do justice to the further range of physical circumstances in which we see redness_e, as discovered by Land and others.[23] The postulate that correlates physical conditions with the *experience* of redness_e will be vastly more complex, for it is reasonable to suppose that all possible *neurological processes* that correlate with this experience are highly complex and diverse, the specification of the physical state of any *single* neuron being a highly complex matter, let alone the specification of *many* neurons, of diverse types, interacting with each other in the somewhat different ways that are, experientially, indistinguishably 'the visual experience

[22] That something along these lines might be possible has been suggested by Nagel: see T. Nagel, 'Conceiving the Impossible and the Mind–Body Problem', *Philosophy* **73** (1998), 337–52. I agree with Nagel when he calls for a conceptual revolution in order to solve the mind–body problem, but disagree with him about the nature of the revolution required. In my view the revolution required is the one argued for in op. cit., note 8. For a discussion of Nagel's suggestion see R. Harré, 'Nagel's Challenge and the Mind–Body Problem', *Philosophy* **74** (1999), 247–70.

[23] E. Land, 'Experiments in color vision', *Scientific American* **200** (1959), 84–99. See also E. Thompson, *Colour Vision* (London: Routledge, 1995).

https://doi.org/10.1017/S003181910000005X Published online by Cambridge University Press

of redness_e'.[24] In addition, it is reasonable to suppose that the list of distinct kinds of experiential features, actual and possible, is all but endless. We might suppose that there are 10^{10} such distinct experiential features. T* will thus consist of 10^{10} postulates in addition to those of T, each postulate being in itself incredibly complex. Whereas T (we are presupposing) is a beautifully unified, explanatory theory, T* is grotesquely complex, disunified and non-explanatory. (And of course almost all of it would be incomprehensible to us in any case; in order to understand all of T*, one would need to have a brain that is made up of all possible conscious brain-structures, stuck together as it were, so that one can oneself experience all possible experiential features of things.)

The upshot of the argument is simply this. In order to develop the beautifully *explanatory* theories that we have in physics, such as Newtonian theory or quantum theory, it is essential that the incredible complexity of the experiential be ignored. This is the price that we pay for being able to explain and understand phenomena physically. If we attempt to develop more complete predictive theories which include extra postulates that link together physical and experiential states of affairs, such theories inevitably become hopelessly non-explanatory.[25]

9 Expanding Personalistic Explanation

If expanding scientific explanation cannot render the experiential intelligible within the physical, could expanding personalistic explanation achieve this? If all actual and possible conscious or sentient brains are taken into account, then there is, we may suppose, a vast realm of the experiential; we human beings are aware only of a minute fragment of this universe of possible experience. Could we imagine a God-like brain that accommodates within itself all possible conscious or sentient brain-structures, so that this God-like mind is able to experience everything that any conscious being

[24] Some philosophers, notably Hilary Putnam, have denied that there are laws correlating the physical and mental aspects of processes going on in the brain. What this denial comes down to is that there are no *simple* laws correlating physical and mental aspects of brain processes. Such laws would have to list the many different kinds of brain processes that correlate with each mental process, such as the visual sensation of redness. But this just reinforces my point that T* would be a horribly complex, and therefore non-explanatory theory.

[25] This argument requires that we can say what it is that makes one theory 'simple', 'unified' or 'explanatory', another theory 'complex', 'disunified', 'non-explanatory': for this see op. cit,. note 1, chapters 3 and 4.

https://doi.org/10.1017/S003181910000005X Published online by Cambridge University Press

whatsoever can experience? Might this God-like mind be able to discern an order, an underlying unity, in the experiential realm, that is for us forever a closed book, simply because of our very limited brains, in comparison?

We might even imagine that this God-like being is able to entertain a single, supreme Idea, which contains within itself everything that any sentient or conscious being can experience, think, decide, desire, feel. This single Idea would correspond, in the experiential world, to the unified unchanging something postulated to exist in the physical universe by the yet-to-be-discovered true, unified theory-of-everything of theoretical physics. Having entertained this supreme unifying Idea, the God-like mind would be able to discern intelligibility, underlying unity in the experiential realm, whereas we, with our vastly more restricted experience, can only discern disunity, inexplicable variety, disorder. The God-like being would understand why such apparently utterly diverse experiences as experience of colour, smell, sound, touch, pain and pleasure exist and are merely just understandable variants of the one, single, supreme, unifying Idea.

It is not easy to see how this fantastic suggestion is to be effectively criticized, once the basic point is conceded that we are debarred from ever knowing or understanding what the great unifying Idea is. *Of course* it seems to us unimaginable or inconceivable that such a unifying Idea should be possible: just that is built into the suggestion!

Some critical comments are, however, possible. It may be doubted that the God-like brain is a physical possibility: it would be impossible to get nutrients, or power, to the brain in question; 'neurons' would not signal sufficiently rapidly. Even if this objection is waived (perhaps because all that is required is the *logical* possibility of the God-like brain, not the *physical* possibility), there may, nevertheless, be doubts about whether it would be functionally possible to have functionally quite different brains accommodated in the one brain. There are, of course, horrendous intellectual and moral problems that lie in wait of any human effort to explore this unknown experiential universe. Are there going to be human being volunteers prepared to undergo brain surgery to have new brain structures built onto existing human brains? If robot technology develops to such an extent that robots can be built capable of having a vastly increased range of experiences, could it conceivably be moral to build such conscious beings, prey, possibly, to nightmarish experiences of which we know nothing? And even if all these objections are waved aside, and beings are built that have a range of experience vastly more extensive than ours, the rest of us would still

https://doi.org/10.1017/S003181910000005X Published online by Cambridge University Press

remain for ever in the dark as to what it is that these beings have learnt.

And there is another, important point. Even if the God-like being exists, and entertains the great, unifying Idea, thus being able to understand a supreme personalistic explanation for the vast diversity and multiplicity of experience, nevertheless this would still leave the essential mystery of the mind–brain problem intact. Even if the God-like being knew how to correlate personalistic and functionalistic descriptions of brain processes and states, the fundamental mystery would, it seems, remain: Why is *this* experience, the visual sensation of redness, let us say, correlated with *this* functionally-described brain process (whatever it may be)? Or are we to suppose that knowing how to correlate the supreme, unifying Idea with its corresponding functionally-described brain state or process somehow leads to a resolution of this key problem?

If this last suggestion is rejected, the task of *explaining* the correlations in question faces the following severe difficulty. What kind of explanation is to be employed for the task? We have seen that scientific (or physical) explanation cannot be employed, and no expansion of scientific explanation can succeed. Personalistic explanation may presuppose the intelligibility of such basic items of experience as the visual sensation of redness but does not explain correlations between experiences and functionally-described brain processes. If we assume that no expansion of personalistic explanation would do the trick either, we are left without any clear candidate for a kind of explanation capable of rendering the correlations comprehensible.

10 Conclusion

An important part of the mind–brain problem arises because sentience and consciousness seem inherently resistant to scientific explanation and understanding. The solution to this dilemma is to recognize, first, that scientific explanation can only render comprehensible a selected aspect of what there is, and second, that there is a mode of explanation and understanding, the personalistic, quite different from, but just as viable as, scientific explanation. In order to understand the mental aspect of brain processes—that aspect we know about as a result of having relevant neurological processes occur in our own brain—we need to avail ourselves of personalistic explanation, irreducible to scientific explanation. The problem of explaining and understanding why experiential or mental aspects of brain processes or things should be correlated with certain physical processes, things or states of affairs is a non-problem because there

https://doi.org/10.1017/S003181910000005X Published online by Cambridge University Press

is no kind of explanation possible in terms of which an explanation could be couched. A physical theory, amplified to include the experiential, might be predictive but would, necessarily, cease to be explanatory; and an amplified personalistic explanation could not succeed either. There is, in short, an *explanation* as to why there cannot be an explanation of correlations between physical and mental aspects of processes going on inside our heads.

This conclusion may seem merely negative: in the nature of things, there is no solution to the mind–brain problem. I have three points to make, however, in support of the claim that there is a positive dimension to this proposed solution to the mind–brain problem.

1. The above does not merely deny that mental aspects of brain processes can be explained and understood scientifically. It stresses that mental aspects can be genuinely explained and understood: but personalistically, not scientifically. And, as I have just said, it provides an *explanation* as to why there can be no explanation of correlations between (functionally described) brain processes and inner experiences. In order to establish that there *is* a problem, it is necessary to indicate a *kind* of explanation in terms of which the correlations could, conceivably, be explained.

2. Even if there is no general *explanation* as to why physical and experiential features are correlated in the way that they are, there are, nevertheless, profoundly important, as yet unsolved but solvable problems of knowledge and understanding concerning such correlations. The central, serious task for research is to discover how the two explanatory accounts of what goes on inside our heads, physical and personal, are inter-related. In order to facilitate this task we need to develop a number of intermediate explanatory accounts, so that we have something like the following: (1) physical (2) molecular (3) chemical (4) neurological (5) functional, or in terms of control architecture (6) purposive (7) personalistic. The problem is to discover how these are inter-related, (1) with (2), (2) with (3), and so on. The major problems lie in discovering how (4), (5), (6) and (7) are inter-related.

A few words about (6) and (5), to take them in reverse order. In addition to personalistic explanations, we need to consider type (6) purposive explanations, which render intelligible the actions of a goal-pursuing thing, whether plant, animal, person, robot or thermostat, by explaining the actions as being designed to realize the overall goal in the given environment, but without appealing in any way to sentience or consciousness. In this respect, purposive explanations are degenerate personalistic explanations, devoid of the element of enabling one to know what it would be oneself to be that

https://doi.org/10.1017/S003181910000005X Published online by Cambridge University Press

robot, oak tree, thermostat, or whatever. Functionalist, or control explanations, in turn, specify control mechanisms, feedback mechanisms and so on, which enable a purposive thing to pursue its goals more or less successfully in the given environment.

In tackling the problem of how explanatory descriptions of head processes are inter-linked, a major task is simply to develop, to create, explanatory descriptions of type (5) that are sufficiently rich and contentful, sufficiently sophisticated, to accommodate the extraordinarily rich and sophisticated control architecture of a conscious human brain. Current explanatory tools may be as inadequate as, let us say, tools of explanation available to Galileo would be were one to attempt to use them to formulate quantum theory and general relativity.

Nevertheless, a part of what needs to be done in an attempt to develop more adequate type (5) explanatory accounts of conscious human brains is to put forward rival conjectures as to how control-correlates of consciousness control, or partly control, human action, all the time seeking to interconnect such control explanatory conjectures with type (4) neurological explanations, on the one hand, and type (6) purposive explanations, on the other hand. Rival conjectures of this type are needed which have the added bonus of being *testable*. If two or more rival, plausible conjectures of this type can be put forward such that, subsequent empirical research confirms one and refutes the rivals, then this branch of psychoneurology would have reached the stage that cosmology reached when it became a part of science through the empirical refutation of the steady state theory, and the confirmation of the big bang theory. In developing such conjectures we need to be guided both by what we know about brain function, and what we know about ourselves as a result of personalistic explanation of human action.

There is another substantial, key problem to be tackled, and another, related methodological path to be followed in tackling the serious, solvable problems just indicated. The problem is: Granted that we understand how it is possible for there to be beings, such as ourselves, open to be explained and understood in the above two very different ways, physical and personalistic[26] (or in the above seven different ways), what explanation is there for the miracle of such doubly comprehensible beings actually existing naturally? This problem was solved, in essence, by Darwin. What neo-Darwinianism does is provide a blind (purposiveless) *mechanism* for the generation of purposive things, living things that pursue the

[26] The problem of how *purposive* beings in the physical universe are possible is solved by pointing to any feedback mechanism, such as a thermostat.

https://doi.org/10.1017/S003181910000005X Published online by Cambridge University Press

goals of survival and reproductive success in ever more diverse and complex ways.

From this Darwinian standpoint, the function of the brain is clear. It is so to control, or guide, the animal so that it acts in its given environment in such ways as to be conducive to survival and reproductive success. Brains have been designed by evolution to be control systems guiding animals to pursue survival and reproductive success. To the above seven types of explanation we need to add an eighth, namely: (8) historical explanation, in particular the historical explanation of neo-Darwinianism.

This eighth kind of explanation presupposes, and uses, some (or, where relevant, all) of the other seven. A basic task of evolutionary biology is to explain how beings have gradually evolved that are simultaneously comprehensible in two or more ways. But because type (7) explanations are not reducible to type (6), (5), … or (1) explanations, evolutionary explanations without personalistic explanations cannot, of themselves, explain the emergence, the evolution, of consciousness.[27]

The methodological path mentioned above is simply this. In tackling the above serious mind–brain problem of discovering how type (4) to (7) explanatory descriptions of head processes are interrelated, it is important to try to retrace the path of evolution. The first step is to solve the serious 'mind–brain' problem for organisms with the simplest possible nervous systems. When this has been accomplished, increasingly complex brains and ways of life can be progressively tackled, ending up with the most complex of all: human beings.

3. The dualism-of-explanation view that I have argued for in this paper has important implications for our understanding of Darwinian theory, and for the problem of discovering how there can be free will, in a worthwhile sense, in a physically comprehensible universe. I hope to explore these implications in a further paper.

University of London

Acknowledgement

This paper was extensively rewritten while I was a Visiting Scholar at the Center for Philosophy of Science at the University of Pittsburgh, a visit that was funded by the British Academy: I would

[27] For suggestions as to how a 'generalized Darwinian research programme' might be developed, incorporating personalistic understanding, see op. cit., note 8, 267–75.

https://doi.org/10.1017/S003181910000005X Published online by Cambridge University Press

like to thank those who made the visit possible. I would also like to thank John McDowell, Richard Gale, Rick Grush and Peter Machamer, all at the University of Pittsburgh, for stimulating discussion in connection with the paper.

https://doi.org/10.1017/S003181910000005X Published online by Cambridge University Press