What is our Real Knowledge about the Human Being
Pontifical Academy of Sciences, *Scripta Varia* 109, Vatican City 2007
www.pas.va/content/dam/accademia/pdf/sv109/sv109-nagel.pdf

SCIENCE AND THE MIND-BODY PROBLEM

THOMAS NAGEL

The relation of mind to the physical world is something we do not at present understand, except superficially. Pursuit of more fundamental understanding faces difficult questions about reductionism, and about the scope and limits of natural science in its present form.

The modern Mind-Body problem arose out of the scientific revolution of the 17th century. Galileo and Descartes made the crucial conceptual division, by proposing that physical science should provide a mathematical and quantitative description of objective reality (consisting of the primary qualities like shape, size, and motion), while subjective appearances and the secondary qualities like color – how the physical world appears to human perception – were assigned to the mind. It was essential to *leave out* or *subtract* subjective appearances and the human mind from the physical world in order to permit a certain kind of objective spatio-temporal conception of physical reality to develop.

But this exclusion of everything mental from the scope of modern physical science was bound to be challenged eventually. We humans are parts of the world, and the desire for a unified world picture is irrepressible. It seems natural to achieve it by extending the reach of physics and chemistry, in light of their great successes in explaining so much of the natural order. This has been accomplished so far by reduction (to basic elements governed by mathematically expressible laws) followed by reconstruction to show how they combine to yield the complexity we observe. Now it has become clear that our bodies and central nervous systems are parts of the physical world, composed of the same elements as everything else. And molecular biology keeps increasing our knowledge of our own physical composition, operation, and development. Finally, so far as we can tell, our mental lives and those of other creatures, including subjective experiences, are strongly connected with and perhaps strictly

dependent on physical events in our brains and on the physical interaction of our bodies with the rest of the physical world.

What are the options for including all these facts in a single world view? We know that Descartes thought they couldn't be unified. His theory is called Dualism: mind and matter are both real and irreducibly distinct, though they interact. Physical science remains defined by the exclusion of the mental from its subject matter. But there are two familiar ways of unifying mind and matter in a single world picture: roughly, by reducing matter to mind or by reducing mind to matter.

The first strategy dominated European philosophy in the 18th, 19th, and early 20th centuries, under the name of *Idealism*. Mind is the ultimate reality and matter is in some way reducible to it. This attempt to overcome the division from the direction of the mental extends from Berkeley, who rejected the primary-secondary quality distinction and held that physical things are ideas in the mind of God – to the logical positivists, who analyzed the physical world as a construction out of sense data. For reasons I don't fully understand, idealism was largely displaced in later 20th-century analytic philosophy by attempts at unification in the opposite direction, starting from the physical.

Physicalism is the view that only the physical world is irreducibly real, and a place must be found in it for mind, if there is such a thing. This would continue the onward march of physical science, through molecular biology, to full closure by swallowing up the mind in the objective physical reality from which it was initially excluded. The assumption is that physics is philosophically unproblematic, and the main target of opposition is Descartes' dualist picture of 'the ghost in the machine'.

One strategy for making the mental part of the physical world picture is conceptual behaviorism, offered as an analysis of the real nature of mental concepts. This was tried in various versions. Mental phenomena were identified variously with behavior, behavioral dispositions, or forms of behavioral organization. In another version, associated with Wittgenstein and Ryle, mental phenomena were not identified with anything, either physical or nonphysical; instead, mental concepts were explained in terms of their observable behavioral conditions of application – criteria or assertability conditions rather than behavioral truth conditions. All these strategies are essentially verificationist, i.e. they assume that the content of a mental statement consists in what would verify it to an observer. So they reduce mental attributions to the externally observable conditions on the basis of which we attribute mental states to others. If successful, this would

obviously place the mind comfortably in the physical world. And it is certainly true that mental phenomena have behavioral manifestations, which supply our main evidence for them in other creatures.

Yet as analyses, all these theories seem insufficient because they leave out something essential that lies beyond the externally observable grounds for attributing mental states to others, namely the aspect of mental phenomena that is evident from the first-person, inner point of view of the conscious subject: for example the way sugar tastes to you or the way red looks, which seems to be something more than the behavioral responses and discriminatory capacities that these experiences explain. Behaviorism leaves out the inner mental state itself.

In the 1950s an alternative, non-analytic route to physicalism was proposed, one which in a sense acknowledged that the mental was something inside us, of which outwardly observable behavior was merely a manifestation. This was the psycho-physical *Identity Theory*, offered by U.T. Place and J.J.C. Smart not as conceptual analysis but as a scientific hypothesis. It held that mental events are physical events in the brain. $\Psi = \Phi$ (where Ψ is a mental event like a pain or a taste sensation and Φ is the corresponding physical event in the central nervous system). This is not a conceptual truth and cannot be known a priori; it is supposed to be a theoretical identity, like Water = H_2O, which can be confirmed only by the future development of science.

The trouble is that this raises a further question: What is it about Φ that makes it also Ψ? Clearly physicalists won't want to give a dualist answer – i.e. that Φ has a nonphysical property. So defenders of the identity theory tended to be pulled back into different kinds of analytical behaviorism, to analyze in nondualist terms the mental character of brain processes. But this time a causal element was added to the analysis: 'the inner state which typically causes certain behavior and is caused by certain stimuli'. This was required by the need to explain the two distinct references to the same thing that occur in a nonconceptual identity statement. The point is to explain how 'pain' and 'brain state' can refer to the same thing even though they do not mean the same, and to explain this without appealing to anything nonphysical in accounting for the reference of 'pain'. But all these strategies are unsatisfactory for the same old reason: Even with the brain added to the picture, they seem to leave out something essential. (And notice, what they leave out is just what was left out of the physical world by Descartes and Galileo in order to form the modern concept of the physical, namely subjective appearances.)

Another problem was subsequently noticed by Saul Kripke. Identity theorists took as their model for $\Psi = \Phi$ other theoretical identities like Water=H_2O or Heat=Molecular Motion. But those identities, he claimed, are necessary (though not conceptual and not a priori), whereas the Ψ/Φ relation appears to be contingent. This was the basis of Descartes' argument for dualism. He said that since we can clearly conceive of the physical body without the mind, and vice versa, they can't be one thing.

Consider Water=H_2O, a typical scientifically discovered theoretical identity, nonconceptual, at least when first discovered. It means that water is *nothing but* H_2O. You *can't have* H_2O without water, and you don't need anything *more* than H_2O for water. It's water even if there's no one around to see, feel, or taste it. We identify water by its perceptible qualities, but our experiences aren't part of the water. The intrinsic properties of water, its density, liquidity between 0 and 100 centigrade, etc. are all fully explained by H_2O and its properties. The physical properties of H_2O are *logically sufficient* for water.

So if Ψ really is Φ in this sense, and nothing else, then Φ by itself, in its physical properties, should be similarly logically sufficient for the taste of sugar. But it doesn't seem to be. It seems conceivable, for any Φ, that there should be Φ without any experience at all. Experience of taste seems something further, contingently connected with the brain state. And this suggests not identity, but dualism, at least of properties. The same intuition makes it seem conceivable (to you) that I could be a completely unconscious zombie, with no mental life, though behaviorally and physically identical to my actual body.

These various dead ends suggest the Ψ/Φ dualism introduced at the birth of modern science may be harder to get out of than many people have imagined. It has even led some philosophers to eliminativism – the suggestion that mental events, like ghosts and Santa Claus, don't exist at all. But if we don't regard that as an option and still want to find an alternative to dualism, my view is that a unified world picture requires something much more radical than physicalism.

I think we have to reject conceptual reduction of the mental to physical. But the appearance of contingency in their relation may be an illusion. The relation may in fact be a necessary but nonconceptual identity, but it may be concealed from us by the inadequacy of the concepts we now have to describe both Ψ and Φ. Both may be partial descriptions of a deeper underlying reality that manifests itself in these different ways when observed from inside (as a state of oneself) and from outside (as a state of the physical brain). Perhaps there is something we have no conception of, which is

logically sufficient for both Φ and Ψ, and without which there can't be either. This would be a form of Monism (like Spinoza's) that is neither idealist nor materialist.

Most major scientific advances involve the creation of new concepts, postulating unobservable elements of reality that are needed to explain the necessity of natural regularities that appear accidental. The evidence for the existence of such things is precisely that if they existed, they would explain what is otherwise incomprehensible. Certainly the mind-body problem is difficult enough so that we should be suspicious of attempts to solve it with the concepts and methods developed to account for very different kinds of things. Instead, we should expect theoretical progress in this area to require a major conceptual revolution. I believe current physics, chemistry, and molecular biology will not by themselves produce an understanding of how the brain gives rise to the mind. This will require a change at least as radical as relativity theory, the introduction of electromagnetic fields into physics – or the original scientific revolution itself, which can't result in a 'theory of everything', but must be seen as a stage on the way to a more general form of understanding. We ourselves are large-scale complex instances of something both objectively physical from outside and subjectively mental from inside. Perhaps the basis for this identity pervades the world.