Mind and Consciousness: Five Questions*

David J. Chalmers

1. *Why were you initially drawn to philosophy of mind?*

Growing up, I was a mathematics and science geek. I read everything I could in these areas. Every now and then, something would point in a philosophical direction. Perhaps my most important influence was reading Hofstadter's *Gödel, Escher, Bach* as a teenager. I read it initially for the mathematical parts, but it planted a seed for thinking about the mind. Later, Hofstadter and Dennett's *The Mind's I* got me thinking more about the mind–body problem in particular.

At the University of Adelaide, I mainly studied mathematics and computer science. But in my first year, I needed an extra course, so I took one in philosophy. I didn't do very well in the course—in fact, it was the black mark on my undergraduate academic record. I remember thinking that philosophers were very difficult to read. Even Nagel's "What is it Like to be a Bat?", which now seems beautifully clear to me, seemed pretty obscure and full of jargon at that time. But the module in the philosophy of mind, in particular, left an impression on me.

I recall being told that Adelaide was where philosophers first developed the thesis that mental states were brain states. I was skeptical of the historical claim, but it turned out to be more or less right: this was the place where Ullin Place and Jack Smart developed the mind-brain identity theory in the 1950s. I was also skeptical of the philosophical claim: I wanted to believe that mental states were brain states, but I could not see how this could be so, and I was convinced that a much more radical and substantial theory of consciousness would be required to truly make the case.

I didn't formally study more philosophy in Adelaide, but I didn't stop thinking about it. I would talk about the problem of consciousness endlessly with my friends. It seemed even then that this was the most important unsolved problem in science. Every few months I would have a new theory, with my favorite being my patented "theory of abstractions", whose centerpieces were the claims that consciousness is an abstraction, and that every abstraction goes along with

[0]In Patrick Grim, ed. *Mind and Consciousness: Five Questions*. Automatic Press, 2009.

some degree of consciousness. I even presented a brief seminar on this theory to the mathematics department, as it was also part of the theory that numbers were abstractions. This led to the inevitable question: is π conscious? My answer at the time was something to the effect that π is conscious but asleep.

I went to Oxford in 1987 to continue my study of mathematics, but on the way I spent a few months hitchhiking around Europe. A lot of that time was spent by the side of the road writing in a notebook, working through all sorts of philosophical ideas. I thought that when I got to Oxford I would return properly to mathematics, but this didn't happen. I became more and more obsessed with the problem of consciousness, having a few ideas that seemed to me at the time to be breakthroughs. One idea was an argument that zombies, and more generally intelligent creatures without consciousness, were impossible (!), based on the idea that they would inevitably talk about consciousness. Another idea was a development of the abstraction theory into a theory of pattern and information. In retrospect, these ideas seem to me to be interesting although not as important as they seemed to be at the time. Still, the whole process has given me a lot of sympathy for people I often hear from with their own breakthrough ideas about consciousness.

Before long it seemed to me that I owed it to myself to develop these ideas properly, and that I should switch from mathematics to philosophy or perhaps cognitive science. Most of my friends and family thought that this was a crazy idea, as I had a track record in mathematics and no evidence that I would be any good in philosophy. Looking at the situation objectively I had to agree, but from the subjective viewpoint it felt like what I had to do. So I met with various Oxford philosophers of mind (Colin McGinn, Kathy Wilkes), and eventually with Michael Dummett, who was then in charge of graduate admissions, and very keen on getting mathematicians into philosophy. I wrote up papers on the two ideas above, and ended up being admitted to the Oxford graduate program in philosophy.

At the same time, I had doubts. My advisor in mathematics, Michael Atiyah, who was an inspiring figure, returned from some time away and convinced me to give mathematics one more try. I also had the impression that Oxford philosophy was very conservative, and made very little contact with science. In retrospect this impression may have been exaggerated by the fact that a number of the philosophers in my college were Wittgensteinians. So I went back to mathematics for a little while. Not much changed, though. It seemed increasingly to me that contemporary mathematics had moved beyond the era of truly fundamental work, while the study of the mind was an area where the fundamental advances were still to come.

In the middle of 1988, I received a long letter from Doug Hofstadter, to whom I had sent my

articles after writing them. He had liked the articles, and suggested that I move to Indiana to join the research group that he was just setting up there. I had no idea where Indiana was, but I went to visit and found both that it was a pleasant place and that the research group was terrific. So I pulled up sticks and moved.

Doug's research group was a tremendously stimulating environment. It was a house full of graduate students and postdocs in all sorts of areas, talking about every topic imaginable in cognitive science, AI, philosophy, and more. I thought initially that I could pursue my ideas from an AI direction, and I spent a lot of time programming connectionist models and the like. But it eventually became clear that to work on consciousness properly, philosophy was the best way for me to go. So I joined the Ph.D. program in philosophy, and belatedly took courses in the area. There wasn't much philosophy of mind at Indiana (though there was plenty of cognitive science), but I read voraciously in the area, to fill in the background that I lacked. I ended up compiling a huge bibliography in the philosophy of mind for this purpose, which has continued to this day on my website.

I still have large gaps in my philosophical education, especially in the history of philosophy, but overall I am glad to have taken the path that I did. In particular, I am glad that I had a chance to think about various ideas philosophically before I had read much about what the great philosophers had thought. One makes a lot of mistakes and reinvents a lot of wheels, but the thinking process itself is invaluable.

2. *What do you consider your most important contribution to the field?*

The topic I have worked on the most is certainly consciousness. When I entered graduate school at Indiana, at the start of 1989, consciousness was not at all a fashionable topic. I remember being struck by the fact that there were hardly any books on the subject, either in philosophy or in cognitive science. Of course that has changed now! But initially, working on consciousness was a fairly lonely process. Still, I knew this was what I wanted to do: from my perspective, the problem of consciousness was *the* reason why a scientist might move to philosophy, while work on smaller topics in the foundations of cognitive science struck me as interesting diversions.

Coming into philosophy from the outside, it seemed to me that many philosophers were either not taking the problem seriously, seizing on cheap methods to deflate the problem that manifestly weren't up to the job, or were taking it too seriously and placing it outside the boundary of science. I've always thought that we need to acknowledge the problem and then face it head on. So I tried

to do this, in a few articles and in my Ph.D. thesis *Toward a Theory of Consciousness*, which eventually turned into my book *The Conscious Mind*.

The Conscious Mind was much closer to being a traditional work of philosophy than I had envisioned at the start. Along the way, I had become convinced that a rigorous philosophical approach, bringing in tools from the philosophy of language and from metaphysics, was essential at least to getting clear on the foundational issues. Doing this that convinced me, contrary to my initial inclination, that a materialist approach to consciousness cannot succeed. So I became a sort of dualist. But I think of this dualism as growing naturally out of the scientific attitude: one needs to acknowledge (not dismiss) the data, and then come up with theories that are adequate to the challenges that the data pose.

I never really conceived of the book primarily as an argument against materialism. I had initially thought that this foundational material might take just the first chapter or so of the thesis, but it ended up taking up the bulk of the first half, and becoming the part of the book that is the most widely read. Still, some of the more speculative positive ideas survived in the second half of the book: for examples, a chapter on an information-based approach to consciousness is recognizably a descendant of my undergraduate ideas about abstraction. The ideas are put forward pretty tentatively, and I suspect that at the end of the day there may be more promising ways forward. But I hope that this sort of thing has at least encouraged people to think constructively about nonreductive theories of consciousness.

I had always thought that issues about consciousness and the mind–body problem were as important for scientists as for philosophers, and it was important to me to be able to present the central ideas in a way that would interest scientists. In 1994, I got a chance to do this, with the first "Toward a Science of Consciousness" at Tucson. I gave a half-hour talk on the problems of consciousness, starting with a distinction between the "hard" and "easy" problems and then presenting the central elements of a view on which consciousness is fundamental. Something about this caught people's attention, in a manner unlike any talk I've given before or since. This led to all sorts of terrific discussions, and ongoing productive interactions with scientists such as Christof Koch at Caltech and Roger Penrose at Oxford. It also led to my being invited to write an article on these ideas for *Scientific American*, and to a symposium on the ideas in the then-new *Journal of Consciousness Studies*.

It's obvious that much of the impact of these ideas was due to being in the right place at the right time. Scientists and philosophers were just returning to consciousness around this time, and distinguishing easy and hard problems of consciousness simply articulated something that many

or most people recognized already. Certainly there's nothing wildly original about recognizing the problem. If anything, there's just something about the "hard problem" formulation that makes the problem hard to avoid. I like some of the arguments in my papers on this topic, as a way of making the case against reductive views of consciousness without technicality. But I can't really see the distinction between the hard problem and the easy problems, which has had the most influence among scientists, as a major contribution to philosophy.

One byproduct of all this is that a lot more people read my book on consciousness than I ever expected, both inside and outside philosophy. It's not an easy book for nonphilosophers, so I'm always pleased when I hear from nonphilosophers who have read it. I was also pleased that philosophers found a lot to chew on in the book, not just in the broad ideas about consciousness but in the connections to metaphysics and the philosophy of language. Many philosophers were returning to the topics of consciousness and the mind–body problem in the 1990s, so this was an exciting time.

I've also ended up working on quite a few other things, inside and outside the philosophy of mind. In the philosophy of mind I've thought a lot about intentionality, trying to develop a broadly Fregean and internalist account of the contents of thought, by developing some of the ideas from two-dimensional semantics that I used in *The Conscious Mind*. At the same time, in other work I've done with Andy Clark, I've looked at the idea that the cognitive processes can extend into the environment. So I seem to have ended up being both an internalist and an externalist, though I don't think that there's really a contradiction here.

I've always had a strong interest in AI and computation, too. More recently I've returned to computational ideas by thinking about the Matrix, which turned out to shed light on a surprisingly large number of other philosophical ideas, at least for me. To think straight about the Matrix, one has to think straight about the philosophy of mind, the philosophy of language, metaphysics, epistemology, the philosophy of physics, the philosophy of computation, and even ethics and the philosophy of religion. The paper I wrote on the Matrix is pretty close to being my favorite among all the papers I've written, and this is a topic that I'm hoping to return to.

Somewhere along the way, I became a philosophical holist. It seems that almost any area of philosophy is relevant to any other. When I first started in philosophy, I was really interested only in the mind–body problem, and questions about, say, sense and reference seemed to me to be nit-picky semantic questions. But to think properly about the mind–body problem, I had to think about metaphysics, and to think about that properly I had to think about the philosophy of language, and to think about that properly I had to think about epistemology, and so on. So I've

ended up doing a fair amount of work in these areas, to the extent that I have a couple of books on these topics (one on meaning and content, one on foundations) that I hope to finish before too long. One pleasant side-effect of this holism is that it has made almost everything in philosophy seem interesting to me.

It's inevitable that work on more specialized philosophical topics has less direct impact than interdisciplinary work on broad themes. But I think that this sort of work is nevertheless crucial, not least to lay foundations for the sort of broader work above and to put it onto a more rigorous footing. I've also come to find it fascinating in its own right. The ideal, I think, is to pursue both big ideas and specialized details in parallel, always doing one with an eye on the other. I like to misquote Kant on this topic: big ideas without details are empty; details without big ideas are blind.

3. *What is the proper role of philosophy in relation to psychology, artificial intelligence, and the neurosciences?*

I have a complicated attitude to the relationship between philosophy and the cognitive sciences. For a start, I think that there is no firm distinction here. In any area of science, one can ask foundational questions. At a certain point, once the questions are foundational enough, one is doing philosophy. But there is no bright line, and the questions can be asked equally by scientists and by philosophers.

There are all sorts of different roles for philosophers to play in these areas. Most straightforwardly, philosophers can help clarify what scientists are up to, and help to distinguish and understand various important but ambiguous or ill-understood ideas in the science (here I think of work on the different notions of representation or of consciousness). Somewhat more ambitiously, philosophers can engage in the process of figuring out just what follows from various empirical results (here I think of work on blindsight or on change blindness, for example). Of course it is usually scientists who generate the data, but the path from data to theory is a process that often involves quasi-philosophical reasoning, and in which philosophers can play a central role. More ambitiously still, philosophy can offer guidance to the sciences, by pointing to promising avenues to explore (here I think of Fodor on modularity), or by making the case that some strategies are more likely to succeed than others (say, in devising a theory of consciousness). Finally, philosophers can sometimes make direct contributions to the sciences, whether by generating data (as experimental philosophers do), by proposing or proving some theoretical principle (as some

Bayesian philosophers do), or by overturning others.

People sometimes say that philosophers shouldn't be prescriptive toward scientists. I don't really see why this is so. Scientists themselves are certainly often prescriptive toward other scientists, and their prescriptions are often based on foundational considerations. There's no reason in principle why philosophers can't do the same. Of course it's a good idea for philosophers to really know the science when they do this, though. And the prescriptions will usually be conditional: *If* you want to explain X, then doing Y won't be enough, and you'll have to do Z. Of course, as with most prescriptions, there is no guarantee that anyone will listen. And many prescriptions will turn out to be wrong. So I am generally in favor of letting a thousand flowers bloom, in science as well as philosophy. But this doesn't mean that some flowerpatches aren't more promising than others.

Just as interesting is the question of what roles the cognitive sciences have to play in philosophy. Again, I don't think there is any bright line here, and answers to foundational questions can be given equally by scientists and philosophers. Still, it's interesting to see the areas where the cognitive sciences have and haven't had a big impact on philosophy.

Most obviously, there have come to be huge areas—the philosophy of neuroscience, the philosophy of psychology, the philosophy of AI— that simply couldn't exist without the relevant sciences. Closer to traditional philosophy of mind, there are all sorts of foundational issues about how the mind works that have been transformed by work in science. Is the mind modular? Do we think using symbols? Are there unconscious processes? Are cognitive capacities innate? Still, these are questions that one should have expected to be empirical questions all along, albeit questions that require a lot of philosophical analysis.

When it comes to some of the really big central questions in the philosophy of mind, however—the mind/body problem, the nature of intentionality or of perceptual experience, the problems of mental causation and free will—it's not clear that these have been transformed by the cognitive sciences to the same extent. The sciences have certainly had an impact around the edges, not least by helping us to understand many specific processes and disorders involving consciousness, intentionality, perception, agency, and so on. But when it comes to some of the big traditional debates—those between materialism and property dualism about consciousness, between internalism and externalism about intentionality, between various metaphysical theories of perception, between compatibilism and incompatibilism about free will—the impact has been less than one might expect.

Of course some will chalk this up to the resistance of philosophers, and some will take it as evidence that these weren't the important questions in the first place. But I think something more

is going on here. Whenever empirical results are brought to bear on the philosophical questions, the application requires some sort of philosophical premise to serve as a bridge. And in case of the big philosophical questions above, in order for this premise to be strong enough that the data bears directly on the question, the premise is typically so strong that it is almost as contentious as the philosophical views at issue. So disagreements about these philosophical views simply ramify into disagreements over the bridging premise.

This isn't always the case. Sometimes, when empirical results are applied to philosophical questions, the bridging premise is somewhat less antecedently contentious than the views in question. In these cases, one gets a sort of amplification from a less contentious premise to a more powerful conclusion. But in the case of questions such as those above, this sort of amplification is relatively rare.

I take the moral to be that the debates in question may well have a deeply philosophical core, one that is unlikely to be resolved by the straightforward application of empirical results. Instead, the core of the debates may well rest on conceptual, metaphysical, and normative issues that fall largely within the a priori domain. So philosophers should not feel embarrassed at spending a lot of time working in a largely non-empirical mode, as most philosophers do. Often this is the best way to get to the heart of the issue.

Different philosophers have different attitudes here. It is not uncommon for scientifically oriented philosophers to hold that there is something deeply old-fashioned or conservative about a priori philosophy, and that the real action lies in the sciences. Perhaps one's background makes a difference here. If one started in traditional philosophy, science can seem refreshing and liberating. But from my own perspective, starting in science, I moved into philosophy precisely because it seemed to address the big questions than science didn't settle. From this perspective it is no surprise that a priori methods should play a major role. I may well also be influenced by having a background in mathematics, which has made enormous progress using largely a priori methods.

Of course this is not to say that philosophers should ignore science. At the very least, philosophers should make sure that their ideas are at least compatible with scientific results. Thinking about science is also a terrific way to help one's thinking about philosophy, not least in expanding one's imagination. Scientific results can be expected to have a major impact on some philosophical questions, and at least a minor impact on almost all philosophical questions. But scientific results are just one tool among many in the philosopher's arsenal.

4. *Is a science of consciousness possible?*

Yes, I think that a science of consciousness is possible. In fact, I think that quite a few bits of it are actual, in contemporary work on consciousness in neuroscience, psychology, and other areas. I've spent a lot of time trying to help get the infrastructure for a interdisciplinary science of consciousness off the ground, through the "Toward a Science of Consciousness" conferences and through the Association for the Scientific Study of Consciousness, as well as through centres devoted to consciousness at Arizona and ANU. It seems to me that the recent explosion in the science of consciousness is one of the most interesting and important intellectual movements of our time.

There are qualifications, of course. I don't think that a successful science of consciousness can be a wholly reductive science of consciousness, cast in terms of neuroscience or computation alone. Rather, I think it will be a nonreductive science, one that does not try to reduce consciousness to a physical process, but rather studies consciousness in its own right and tries to find connections to brain, behavior, and other cognitive processes. If you look at the contemporary neuroscience and psychology of consciousness, this is just what you find. Any attempts at reduction of consciousness are extremely half-hearted. Instead neuroscience is largely engaged in finding neural *correlates* of consciousness, without making claims about reduction. Psychology studies the connections between conscious processes, unconscious processes, and behavior. This way, a lot of progress has been made.

I think that the science of consciousness also differs from many other sciences in that it gives an essential role to subjective or first-person data. In a way, each of our own conscious experiences provide the primary data that is distinctive to the science of consciousness. We cannot directly observe the experiences of others, so our access to consciousness is largely mediated by introspection (in ourselves) and verbal reports (in others). Assuming that we can take the deliverances of introspection and verbal report at face value – which is by no means always the case—these can be used to build up a store of first-person data about consciousness itself. One can then correlate this data with third-person data about brain and behavior, and attempt to integrate all these data via principles that connect them. I think that the principles of a satisfactory science of consciousness will always make ineliminable reference to subjective experience, though.

Of course it is very early days at the moment. We're greatly limited by what we know about the brain. Brain imaging tells one only so much, and the invasive techniques that can tell one more (such as single-cell recording) are largely limited to non-human animals and the occasional surgical patient. We're also limited by our methods for investigating states of consciousness. The sci-

ence typically uses rough-and-ready introspective reports, but these are extremely coarse-grained. Ideally we would like to be able to use sophisticated and reliable introspective techniques, combined with some sort of rich formal language for expressing and analyzing states of consciousness. We don't have anything like that yet, and it's an open question whether these things are possible. But again, it's early in the day. After making a start on these topics in the nineteenth century, the sciences have only recently been returning to them. The proof will be in the pudding.

Ultimately, the hope is for a set of fundamental principles connecting physical processes and consciousness. If I am right about the metaphysics of consciousness, then these principles will have a status akin to that of fundamental laws in physics. I've speculated a bit about what these principles might be, but any theories that we come up with now will almost certainly be wrong. The science of consciousness probably has a revolution or three to go through before it gets to anything like its destination, and when it gets there, it may be quite unlike anything that we currently imagine.

Given this, though, I think we should be open to all sorts of ideas. I'm always pleased to see scientists and philosophers putting forward positive theories of consciousness. Even when they are wrong, one learns something from the attempt. When I first got into philosophy, I was disappointed by how little positive theorizing there was about consciousness. Philosophers seemed to have the sense that theorizing about consciousness should be left to scientists, while scientists seemed to have the sense that theorizing about consciousness should be led to philosophers. That situation has improved to some extent, but I'd like to see more of it. Of course this work often goes out on a limb, but sometimes one has to go out on limbs to get through the forest.

5. *What are the most important open problems in contemporary philosophy of mind? What are the most promising prospects?*

I take it that all of the most important problems in the philosophy of mind are still open. This applies most obviously to the mind–body problem and its various components, such as the problem of consciousness, the problem of intentionality, and the problem of mental causation. But the same goes for most of the other traditional problems in the area: the problem of other minds, the unity of consciousness, the nature of self-knowledge, the nature of concepts, and so on.

It seems to be in the nature of these problems that they are clarified rather than solved. Or perhaps they are solved to the satisfaction of an individual, or solved to the satisfaction of small groups or small communities. But they are never solved to the satisfaction of the philosophical community

as a whole, for any extended period of time. Perhaps it is reasonable to doubt whether they ever will be. Instead, we might just end up with an increasingly good understanding of the fundamental disagreements on which debates over these issues turn, and with a conditional understanding of what one's theory should look like given one's view on these fundamental disagreements.

Still, one can make progress on these problems, and on a host of smaller problems. Often, philosophical progress comes from focusing in a new area where people had not much previously focused. For example, over the last decade or so, there has been an enormous improvement of our understanding of the relationship between consciousness and intentionality, after years in which the topics were largely treated separately in analytic philosophy. There are increasingly sophisticated analyses of the intentional structure of consciousness, and people are beginning to look at the converse question of the role that consciousness might play in intentionality. There are still rich pickings in this area, and I expect to see a lot more progress in the next decade or so.

On the mind–body problem, it's not surprising that I think that some of the richest pickings will come from developing nonreductive approaches to consciousness in real depth. One approach that is drawing increasing attention is that of Russellian monism: grounding consciousness in the unknown intrinsic properties of matter. The idea is too strange for some philosophers, but we've learned that the world is a strange place. I think that if someone can really take this idea and develop it properly, it has the potential to end up as a truly powerful approach to the problem.

There is also an enormous amount to be learned about specific aspects of consciousness. There have been many advances in the philosophical study of perception and perceptual consciousness in recent years. I think that we may be on the threshold of a period of advances in the study of conscious thought. Other aspects of consciousness that promise to yield a great deal in the coming years include temporal consciousness and the relation between consciousness and attention.

One can certainly expect that there will also be a huge amount of philosophical activity driven by the latest results coming from the cognitive sciences. Neuroscience will attract an increasing amount of attention. Although there is a reasonable amount of philosophy of neuroscience at the moment, it is surprising that there isn't more. I have begun some collaborative work with neuroscientists myself, on the question of detecting consciousness in patients diagnosed with vegetative state and related post-coma conditions. This is a place where neuroscience comes together with the philosophical problem of other minds in interesting ways. I think that philosophers have a lot to contribute to areas like this, and I hope that more philosophers will move in these directions.

Of course philosophy, like other academic disciplines, is subject to vicissitudes of fashion. This hurt the study of consciousness for many years, and more recently has helped it. In parallel,

the study of intentionality saw a huge surge around the 1980s, followed by a swing away from the area as early promises seemed not to pan out. I have the sense that philosophers are ready to return to the study of intentionality, though, perhaps enriched by what we have learned about consciousness in the meantime, as well as by what we have learned in the philosophy of language.

Speaking for myself, I will continue to work on foundational ideas in metaphysics, epistemology, and philosophy of language, trying to pull together the details of a big picture that I hope can be used to shed light on many philosophical questions. At the end of the day, though, I am a philosopher of mind, and the problem of consciousness remains my first love. Before I die, I'd like to have one or two more cracks at coming up with a positive theory of consciousness. My older self says that this is probably quixotic, but my younger self says that one should at least try.