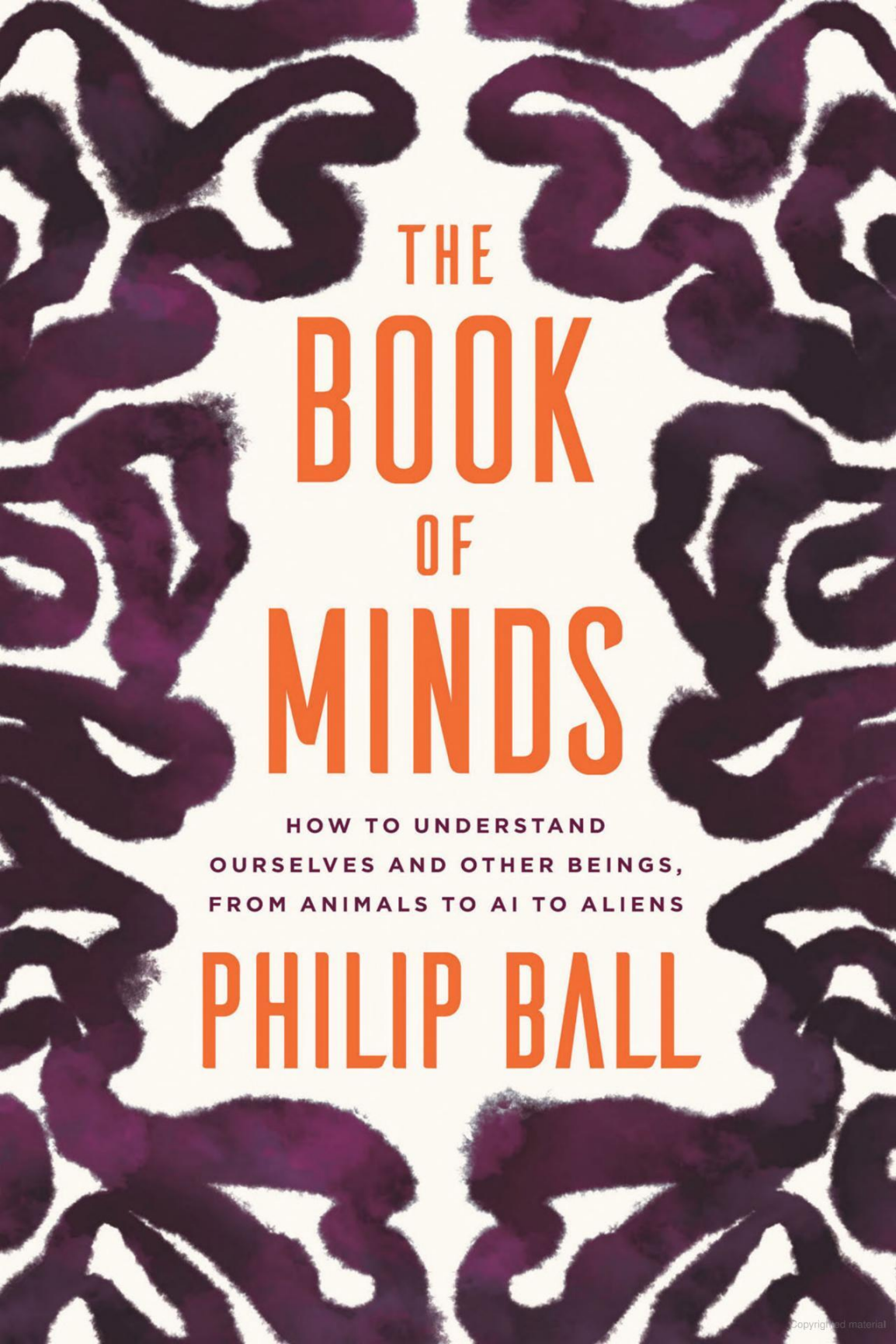

The University of Chicago Press, Chicago 60637

Published 2022
Printed in the United States of America

31 30 29 28 27 26 25 24 23 22 1 2 3 4 5

ISBN-13: 978-0-226-79587-4 (cloth)
ISBN-13: 978-0-226-82204-4 (e-book)
DOI: https://doi.org/10.7208/chicago/9780226822044.001.0001

Published in the United Kingdom by Picador, an imprint of Pan Macmillan.

Library of Congress Cataloging-in-Publication Data

Names: Ball, Philip, 1962– author.
Title: The book of minds : how to understand ourselves and other beings, from animals to AI to aliens / Philip Ball.
Description: Chicago : The University of Chicago Press, 2022. | Includes bibliographical references and index.
Identifiers: LCCN 2021061676 | ISBN 9780226795874 (cloth) | ISBN 9780226822044 (ebook)
Subjects: LCSH: Cognition. | Consciousness. | Brain. | Cognition in animals. | Artificial intelligence.
Classification: LCC BF311 .B27 2022 | DDC 153—dc23/eng/20220114
LC record available at https://lccn.loc.gov/2021061676

This paper meets the requirements of ANSI/NISO Z39.48-1992 (Permanence of Paper).

Contents

CHAPTER 1

Minds and Where to Find Them

The neurologist and writer Oliver Sacks was an indefatigable chronicler of the human mind, and in the too-brief time that I knew him I came to appreciate what that meant. In his personal interactions as much as his elegant case-study essays, Sacks was always seeking the essence of the person: how does *this* mind work, how was it shaped, what does it believe and desire? He was no less forensic and curious about his own mind, which I suspect represented as much of a puzzle to him as did anyone else's.

Even – perhaps especially – for a neurologist, this sensitivity to minds was unusual. Yes, Sacks might consider how someone's temporal lobe had been damaged by illness or injury, and he might wonder about the soup of neurotransmitters sloshing around in the grey matter of the brain. But his primary focus was on the individual, the integrated result of all that neural processing: a person existing, as best they could, in the company of others, each trying to navigate a path amid other minds that they could never really hope to fathom and certainly never to experience. It was emphatically not the brain but the mind that fascinated him.

None more so, I imagine, than the mind he once encountered in Toronto, even though he was never able to make a case study of it. That would not have been easy, because this individual was not

human. In the city zoo, Sacks had this briefest of exchanges with a female orangutan.

> She was nursing a baby – but when I pressed my bearded face against the window of her large, grassy enclosure, she put her infant down gently, came over to the window, and pressed her face, her nose, opposite mine, on the other side of the glass. I suspect my eyes were darting about as I gazed at her face, but I was much more conscious of her eyes. Her bright little eyes – were they orange too? – flicked about, observing my nose, my chin, all the human but also apish features of my face, identifying me (I could not help feeling) as one of her own kind, or at least closely akin. Then she stared into my eyes, and I into hers, like lovers gazing into each other's eyes, with just the pane of glass between us.
>
> I put my left hand against the window, and she immediately put her right hand over mine. Their affinity was obvious – we could both see how similar they were. I found this astounding, wonderful; it gave me an intense feeling of kinship and closeness as I had never had before with any animal. 'See,' her action said, 'my hand, too, is just like yours.' But it was also a greeting, like shaking hands or matching palms in a high five.
>
> Then we pulled our faces away from the glass, and she went back to her baby.
>
> I have had and loved dogs and other animals, but I have never known such an instant, mutual recognition and sense of kinship as I had with this fellow primate.

A sceptic might question Sacks' confident assertion of meaning: his supposition that the orangutan was expressing a greeting and was commenting on their physical similarities, their kinship. You don't know what is going on in an ape's mind! You don't even know that an ape *has* a mind!

But Sacks was making this inference on the same grounds that we infer the existence of *any* other mind: intuitively, by analogy with our own. All we know for sure, as René Descartes famously observed, is that we exist – by virtue of our own consciousness, our sense of *being* a mind. The rest is supposition.

No one has ever got beyond Descartes' philosophical conundrum: how can we be *sure* of anything but ourselves? Imagine, Descartes said, if there were some mischievous demon feeding our mind information that creates the perfect illusion of an external world, filled with other minds like ours. Maybe none of that is real: it is all just phantoms and mirages conjured by the demon as if by some trick of infernal telepathic cinematography.

This position is called solipsism, and widely considered a philosophical dead end. You can't refute it, yet to entertain it offers nothing of any value. If other people are merely a figment of my imagination, I cease to have any moral obligations towards them – but to assume without evidence that this is the case (rather than at least erring on the side of caution) would require that I embrace a belief my experience to date has primed me to regard as psychotic. It would seem to invert the very definition of reason. At any rate, since the rational solipsist can never be sure of her belief, it can't necessitate any change in her behaviour: it's an impotent idea. Sure, the demon (and where then did *he* come from?) might decide to end the game at any moment. But everything that transpires in my mind advises me to assume he will not (because I should assume he does not exist).

We'll hear more from Descartes' demon later, because scientific and technological advances since the seventeenth century have produced new manifestations of this troublesome imp. Let us return to the mind of Oliver Sacks's orangutan.

What, exactly, persuaded Sacks that the ape had a kindred mind? The familiar gestures and the soulful gaze of apes seem to insist on

it. This intuition goes beyond anatomical similarities; apes, probably more than any other creatures (dog owners might demur), exhibit a deeply eloquent quality of eye contact. Those eyes are too expressively reminiscent of what we see in other humans for us to imagine that they are the windows to vastly different minds – let alone to no mind at all. In short, there is a great deal we import from encounters with other people into meetings with our more distant primate relatives.

It's the similarity of our own behaviour to that of other people which convinces us they too have 'somebody at home': that a mind like ours inheres within the other's body, guiding its outward actions. It is simply more parsimonious to suppose that another person is a being just like us than to imagine that somehow the world is peopled with zombie-like beings able, through some bizarre quirk of physics or biology, to mimic us so perfectly. How weird and improbable it would be if the inscrutable laws of zombiehood impelled these other beings to use language (say) in just the same way as we do, yet without any of the same intent. What's more, other people's brains produce patterns of electrical activity identical in broad outline to those in our own, and the same patterns correlate with the same behaviours. Such coincidences, if that is all they were, could surely only be the product of demonic design.

Far from being a leap of faith, then, assuming the existence of other minds is the rational thing to do – not just in people but also in orangutans. We'll see later that this argument for the reality of animal minds – the assumption that they are not merely complex automata, rather as Descartes supposed – can be made much more concrete.

But how far can this reasoning take us? I hope you're willing to grant me a mind – I can assure you (though of course you have only my word for it) that you'd be doing the right thing. I suspect most people are now happy, even eager, to accept that it is meaningful to

say that apes have minds. The difficulty in going further, however, is not that we are habitually reluctant to attribute minds to non-human beings and entities, but that we do this all too readily. We have evolved to see minds every damned place we look. So some caution is in order.

Is this world not so glorious and terrible in its profuse and sublime extent that it must constitute evidence of a minded* Creator? That's what humankind has long believed, filling all corners of the world with entities that possess mind and motive. That wind? The spirits of the air are on the move. That crash of thunder? The storm god is restless. That creaking floorboard? The tread of a ghostly being.

It has become common in our increasingly secular age to treat all this animism either as a quirk of our evolutionary past that we need to outgrow or, worse, as evidence that we are still in thrall to pre-scientific delusions. I suggest our tendency to attribute mind to matter is a lot more complicated than that. What if, for instance, in stripping the world of mindedness we sacrifice our respect for it too, so that a river devoid of any animating spirit will eventually become no more than a resource to be exploited and abused? As we will see, some scientists today seriously argue that plants have minds, partly on the grounds that this should deepen our ecological sensitivity. It is surely no coincidence that the British scientist and inventor James Lovelock, having conceived of the entire Earth as a self-regulating entity with organism-like properties, accepted the suggestion of his neighbour, novelist and Nobel laureate William Golding, to personify that image with the name of the Greek earth goddess Gaia. For some, these ideas veer too far from science and too close to mysticism. But the point is that the impulse to award mindedness

* I will be using the word *minded* in a somewhat unconventional sense, to mean *imbued with mind*.

where it does not *obviously* reside – and thereby to valorize the minded entity – has not gone away, and we might want to consider if there are good reasons why that is so.

I doubt that even the hardest-headed sceptic of our instinct to personify nature, objects, and forces has not occasionally cursed the sheer perversity or bloody-mindedness of their computer or car. 'Don't do that!' we wail in futile command as the computer decides to shut down or mysteriously junks our file. (I feel that right now I am tempting fate, or the Computer God, even to say such a thing.) 'Why me?' we cry when misfortune befalls us, betraying the suspicion that deep down there is a reason, a plan, that the universe harbours. (That, in a nutshell, is the Book of Job, which in a generous reading warns against interpreting another's bad luck as an indication that God, displeased with them, has meted out their just deserts.)

If there's a flaw in our tendency too casually to attribute mind, it might better be located in the anthropocentric nature of that impulse. We can't resist awarding things *minds like ours*. As we'll see later, Christian theologians have striven in vain to save God from that fate, and it is no wonder: their own holy book undermines them repeatedly, unless it is read with great subtlety. (God seems to speak more regularly to Noah, Jacob and Moses than many company CEOs today do to their underlings.) Our habit of treating animals as though they are dim-witted humans explains a great deal about our disregard for their well-being; giving them fully fledged, Disneyfied human minds is only the flipside of the same coin. We'll see too that much of the discussion about the perils of artificial intelligence has been distorted by our insistence on giving machines (in our imaginations) minds like ours.

Still, it's understandable. As we are reminded daily, it's hard enough sometimes to fathom the minds of our fellow humans – to accept that they might think differently from us – let alone to

imagine what a non-human mind could possibly be like. But that is what we're going to try to do here, and I believe the task is not hopeless.

Making minds up

First of all, we need to ask the central question: *What is a mind?*

There is no scientific definition to help us. Neither can dictionaries, since they tend to define the mind *only* in relation to the human: it is, for example, 'the part of a person that makes it possible for him or her to think, feel emotions, and understand things.' It's bad enough that such a definition leans on a slew of other ill-defined concepts – thinking, feeling, understanding. Worse, the definition positively excludes the possibility of mind existing within non-human entities.

Some behavioural researchers dislike the word altogether. 'Mind', say psychologists Alexandra Schnell and Giorgio Vallortigara, 'is an immeasurable concept that is not amenable to rigorous scientific testing.' They say that instead of talking about the 'dog mind' or the 'octopus mind', we should focus on investigating the mechanisms of their cognition in ways that can be measured and tested.

They have a point, but science needs vague concepts as well as precise ones. 'Life' too is immeasurable and untestable – there is no unique way to define it – and yet it is an indispensable notion for making sense of the world. Even words like 'time', 'energy', and 'molecule' in the so-called hard physical sciences turn out to be far from easy to define rigorously. Yet they are useful. So can the idea of mind be, if we are careful how we use it.

It's understandable yet unfortunate that much of the vast literature on the philosophy of mind considers it unnecessary to define its terms. Gilbert Ryle's influential book *The Concept of Mind* (1949) is so confident that we are all on the same page from the outset that

it plunges straight into a discussion of the attributes that people display. Daniel Dennett, one of the most eloquent and perceptive contemporary philosophers of mind, presents a nuanced exploration of what non-human minds might be like in his 1996 book *Kinds of Mind*, and yet he too has to begin by suggesting that, 'Whatever else a mind is, it is supposed to be something like our minds; otherwise we wouldn't call it a mind. So our minds, the only minds *we* know from the outset, are the standard with which we must begin.'

He is right, of course. But this constraint is perhaps only because, in exploring the *Space of Possible Minds*, we are currently no better placed than the pre-Copernican astronomers who installed the Earth at the centre of the cosmos and arranged everything else in relation to it, spatially and materially. Our own mind has certain properties, and it makes sense to ask whether other minds have more or less of those properties: how close or distant they are from ours. But this doesn't get us far in pinning down what the notion I am referring to as *mindedness* – possessing a mind – means. One thing my mind has, for example, is memory. But my computer has much more of that, at least in the sense of holding vast amounts of information that can be recalled exactly and in an instant. Does that mean my computer exceeds me in at least this one feature of mind? Or is memory in fact not a necessary requirement of *mindedness* at all? (I shall answer this question later, after a fashion.)

In short, 'mind' is one of those concepts – like intelligence, thought, and life – that sounds technical (and thus definable) but is in fact colloquial and irreducibly fuzzy. Beyond our own mind (and what we infer thereby about those of our fellow humans), we can't say for sure what mind should or should not mean. We are not really much better off than what Ambrose Bierce implied in his satirical classic of 1906, *The Devil's Dictionary*, where he defined mind as

> A mysterious form of matter secreted by the brain. Its chief activity consists in the endeavor to ascertain its own nature, the futility of the attempt being due to the fact that it has nothing but itself to know itself with.

Yet I don't believe that a definition of mind need be impossible, so long as we're not trying to formulate it with scientific rigour. On the contrary, it can be given rather succinctly:

For an entity to have a mind, there must be something it is like to be that entity.

I apologize that this is a syntactically odd sentence, and therefore not easy to parse. But what it is basically saying is that a mind hosts an experience of some sort.

Some might say this is not a properly scientific definition because it invokes subjectivity, which is not a thing one can measure. I'm agnostic about such suggestions, both because there *are* scientific studies that aim to measure subjective experience and because a concept (like life) doesn't have to be measurable to be scientifically useful.

You might, on the other hand, be inclined to object that this definition of mind is tautological. What else could a mind be, after all?* But I think there is a very good reason for making it our starting point, which is this: the only mind we know about is our own, and *that* has experience. We don't know *why* it has experience, but

* All the same, other definitions exist. For example, neuroscientists Ogi Ogas and Sai Gaddam require of a mind only that it 'takes a set of inputs from its environment and transforms them into a set of environment-impacting outputs that influence the welfare of its body.' It's not hard to make machines that do more or less this – and indeed Ogas and Gaddam seem to consider machine minds to be an unproblematic notion. We will see in Chapter 7 how far my own definition can be extended to machines.

only that it does. We don't even know quite how to characterize experience, but only that we possess it. All we can do in trying to understand mind is to move cautiously outwards, to see what aspects of our experience we might feel able (taking great care) to generalize. In this sense, trying to understand mind is not like trying to understand anything else. For everything else, we use our mind and experience as tools for understanding. But here we are forced to turn those tools on themselves.

That's why there is something irreducibly phenomenological about the study of mind, in the sense invoked by the philosophical movement known as Phenomenology pioneered by Edmund Husserl in the early twentieth century. This tradition grapples with experience from a first-person perspective, abandoning science's characteristic impulse of seeking understanding from an impersonal, objective position. My criterion of mind is, I'd argue, not tautological but closer to phenomenological, and necessarily so when *mind* is the subject matter.

Since we can't be sure about the nature of other minds, we have to be humble in our pronouncements. I do not believe that a rock has a mind, because I don't think a rock has experience: it does not mean anything to say that 'being like a rock' is to be like anything at all. This, however, is an opinion. Some philosophers, and some scientists too, argue that there *is* something it is like to be a rock – even if that is only the faintest glimmer of 'being like'. This position is called panpsychism:* the idea that qualities of mind pervade all matter to some degree. It could be right, but panpsychists can't prove it.

* This word has sometimes been used with a derogatory implication, as if the idea it denotes is self-evidently absurd. It is not, and indeed the panpsychist position has enjoyed something of a recent resurgence, partly for reasons that we shall discover later.

Yet we need not be entirely mired in relativism. Scientists and philosophers who suspect there might be something it is like to be a rock don't say so because of some vague intuition, or because they cleave to an animistic faith. The claim is one arrived at by reasoning, and at least some of that reasoning can be examined systematically and perhaps even experimentally. We are not totally in the dark.

How about a bacterium? Is there something it is like to be a bacterium? Here opinions are more divided. Some invoke the notion of *biopsychism*: the proposal that mindedness is one of the defining, inevitable properties of all living things. We'll look at this position more closely later, but let's allow for now that it is not obviously crazy. Personally, I'm not sure I believe there is something it is like to be a bacterium.

Still, you can see the point. At some stage on the complexity scale of life, there appears some entity for which there is something it is to be like that organism. I imagine most people are ready to accept today that there is something it is like to be an orangutan. You might well consider there is something it is like to be a mouse, perhaps even a fly. But a fungus? Maybe that's pushing it.

This is why it makes sense to speak in terms of mindedness, which acknowledges that minds are not all-or-nothing entities but matters of degree. My definition notwithstanding, I don't think it's helpful to ask if something has a mind or not, but rather, to ask what qualities of mind it has, and how much of them (if any at all).

You might wonder: why not speak instead of consciousness? The two concepts are evidently related, but they are not synonymous. For one thing, consciousness seems closer to a property we can identify and perhaps even quantify. We know that consciousness can come and go from our brains – general anaesthesia extinguishes it temporarily. But when we lose consciousness, have we lost our mind too? It's significant that we don't typically speak of it in those terms. As we will see, there are now techniques for measuring

whether a human brain is conscious; they are somewhat controversial and it's not entirely clear what proxy for consciousness they are probing, but they evidently measure *something* meaningful. What's more, even though we still lack a scientific theory of consciousness (and might never have such a thing), there is a fair amount we can say, and more we can usefully speculate, about how consciousness arises from the activity of our neurons and neural circuits.

Being minded, on the other hand, is a capacity that is both more general and more abstract: you might regard the condition as one that entails being conscious at least some of the time, but that more specifically supplies a repertoire of ways to feel, to act and simply to be.

We might say, then, that mindedness is a disposition of cognitive systems that can potentially give rise to states of consciousness. I say *states* because it is by no means clear, and I think unlikely, that what we call consciousness corresponds to a single state (of mind or brain). By the same token I would suggest that while there's a rough-and-ready truth to the suggestion that greater degrees of mindedness will support greater degrees of consciousness, neither attribute seems likely to be measurable in ways that can be expressed with a single number, and in fact both are more akin to qualities than quantities. Different types of mind can be expected to support different kinds of consciousness. Can mind exist without any kind of consciousness at all? It's hard to imagine what it could mean to 'be like' an entity that lacks any kind of awareness – but as we'll see, we might make more progress by breaking the question down into its components.

Colloquial language is revealing of how we think about these matters. 'Losing one's mind' implies something quite different to losing consciousness; here what is really lost is not the mind per se but the kind of mind that can make good (beneficial) use of its resources. Mind is a verb too, implying a sort of predisposition:

Mind out, mind yourself, would you mind awfully, I really don't mind. We seem to regard mind as disembodied: mind over matter, the power of the mind. As we'll see, there is probably on the contrary a close and indissoluble connection between mind and the physical structure in which it arises – but the popular conception of mind brings it adjacent to the notions of will and self-determination: a mind does things, it achieves goals, and does so in ways that we conceptualize non-physically.

By what means does a mind enact this functional objective? Philosopher Ned Block has proposed that the mind is the 'software of the brain' – it is, you might say, the algorithm that the brain runs to do what it does. He identifies at least two components to this capability: intelligence and intentionality. Intelligence comes from applying rules to data: the mind-system takes in information, and the rules turn it into output signals, for example to guide behaviour. That process might be extremely complex, but Block suggests that it can be broken down into progressively less 'intelligent' subsystems, until ultimately we get to 'primitive processors' that simply convert one signal into another with no intelligence at all. These could be electronic logic gates in a computer, made from silicon transistors, or they could be individual neurons sending electrical signals to one another. No one (well, hardly anyone) argues that individual neurons are intelligent. In other words, this view of intelligence is agnostic about the hardware: you could construct the primitive processors, say, from ping pong balls rolling along tubes.

But intelligence alone doesn't make a mind. For that, suggests Block, you also need intentionality – put crudely, what the processors involved in intelligence are *for*. Intentionality is *aboutness*: an intentional system has states that in some sense represent – are about – the world. If you stick together a strip of copper and one of tin and warm them up, the two metals expand at different rates, causing the double strip to bend. If you now make this a

component in an electronic circuit so that the bending breaks the circuit, you have a thermostat, and the double strip becomes an intentional system: it is *about* controlling the temperature in the environment. Evidently, intentionality isn't a question of what the system looks like or what, of itself, it does – but about how it relates to the world in which it is embedded.*

This is a very mechanical and computational view of the mind. There's nothing in Block's formulation that ties the notion of mind to any biological embodiment: minds, you might say, don't have to be 'alive' in the usual sense.†

We're left, then, with a choice of defining minds in terms of either their nature (they have sentience, a 'what it is to be like') or their purpose (they have goals). These needn't be incompatible, for one of the tantalizing questions about types of mind is whether it is possible even to conceive of a sentient entity that does *not* recognize goals – or conversely, whether the origin of a 'what it is to be like' resides in the value of such experiential knowledge for attaining a mind's goals. Dennett suggests their key objective by quoting the French poet Paul Valéry: 'the task of a mind is to produce future.' That is to say, Dennett continues, a mind must be a generator of expectations and predictions:

> it mines the present for clues, which it refines with the help of the materials it has saved from the past, turning them into anticipations of the future. And then it acts, rationally, on the basis of those hard-won anticipations.

* Does this mean a thermostat has a mind? Philosophers have in fact debated this question, but have not reached a consensus.

† The question of what constitutes 'being alive' is not settled either, but that's another matter.

If this formulation is correct, we might expect minds to have certain features: memories, internal models of the world, a capacity to act, and perhaps 'feelings' to motivate that action. A mind so endowed would be able not only to construct possible futures, but also to make selections and try to realize them.

Dennett's prescription imposes a requirement on the *speed* with which a mind deliberates: namely, that must happen at a rate at least comparable to that at which significant change happens in the environment around it. If the mind's predictions arrive too late to make a difference, the mind can't do its job – and so it has no value, no reason to exist. Perhaps, Dennett speculates, this creates constraints on what we can *perceive* as mind, based on what we perceive as salient change. 'If', he says,

> our planet were visited by Martians who thought the same sort of thoughts as we do but thousands or millions of times faster than we do, we would seem to them to be about as stupid as trees, and they would be inclined to scoff at the hypothesis that we had minds . . . In order for us to see things as mindful, they have to happen at the right pace.

It's unlikely, as we'll see, that we are overlooking a tree mind simply because it works at so glacial a pace; but Tolkien's fictional Ents serve to suggest that relative slowness of mind need not imply its absence, or indeed an absence of wisdom. Or to put it another way: mindedness might have an associated timescale, outside of which it ceases to be relevant.*

* This illustrates one respect in which our technical devices effectively expand our range of mind. A hundred years ago it made no difference to electrons moving in atoms on attosecond timescales (10^{-21} seconds) whether we had minds or not. But today we can use ultrafast laser pulses to alter those motions intentionally and

Block's view would seem to make mind a very general biological property. If intelligence is a matter of possessing some information-processing capacity that turns a stimulus into a behaviour, while intentionality supplies the purpose and motive for that behaviour by relating it to the world, then all living things from bacteria to bats to bank managers might be argued to have minds.

Neuroscientist Antonio Damasio demands more from a mind. Organisms, and even brains, he says, 'can have many intervening steps in the circuits mediating between response and stimulus, and still have no mind, if they do not meet an essential condition: the ability to deploy images internally and to order those images in a process called thought.'

Here, Damasio does not necessarily mean visual images (although they could be); evidently it is not necessary to possess vision at all in order to have a mind. The imagery could be formed from sound sensations, or touch or smell, say. The point is that the minded being uses those primitive inputs to construct some sort of internal picture of the world, and act on it. Action, says Damasio, is crucial: 'No organism seems to have mind but no action.' By the same token, he adds, there are organisms that have 'intelligent actions but no mind' – because they lack these internal representations through which action is guided. (This depends on what qualifies as a representation, of course.)

But there's still some postponing of the question in this formulation. It teeters on the brink of circularity: a mind is only a mind if it thinks, and thinking is what minds do. It is possible already to build machines that seem to satisfy all of Damasio's criteria – they can, for example, construct models of their environment based on input data, run simulations of these models to predict the

---

with design: our minds can touch and impose their plans on processes that happen far faster than thought itself.

consequences of different behavioural choices, and select the best. This can all be automated. And yet no one considers that these artificial devices warrant being admitted to the club of minded entities, because we have absolutely no reason to think that there is any awareness involved in the process. There is still nothing that it is to be like these machines.

At least, that's what nearly all experts in AI will say, and I believe they are right. But it's not obvious how we could find out for sure. We can, in principle, know everything there is about, say, a bird brain, except for what it is like to be 'inside' it. My definition of mind therefore can't obviously be tested, verified or falsified, any more than can the scenario posed by Descartes' demon. And by the same token, it's not productive to fret too much about that. Rather than arguing over the question of whether other minds exist or not, we can more usefully ask: how does mindedness arise from the cognitive workings of our own brains? Which if any of these cognitive properties are indispensable for that to happen? How might these appear or differ in other entities that might conceivably be minded, and what might the resulting minds be like from the inside? If we have answers, can we design new kinds of mind? Will we? Should we?

Why minds?

Damasio's description of mind is incomplete in a useful way. For if an intelligent system is able to acquire all of these features and yet *still not be a mind*, why is anything more needed, or of any value? Given those capacities, why is it necessary for there to be a 'what it feels like' at all? It's not obviously impossible that our distant evolutionary ancestors evolved all the way to being Damasio's intelligent yet mindless beings, and then natural selection 'discovered' that there was some added advantage to be had by installing a mind

amidst it all. This used to be a common view: that we humans are unique as beasts with mind and awareness, distinguished by the fact that we are not automata but willed beings. It can be found in Aristotle's categorization of living things as those with only a nutritive soul (like plants), those with also a sensitive soul (like animals), and those with also a rational soul or *nous* (us). This exceptionalism persisted in Descartes' claim that humankind alone possesses a soul in the Christian sense: an immortal essence of being. It's not clear how deeply Descartes was persuaded of that, however: his account of the human body presented it, in the spirit of his times, as a machine, a contraption of pumps, levers and hydraulics. He may have insisted on the soul as the animating principle partly to avoid charges of heresy in presenting in so mechanical a manner the divine creation that is humanity. (It didn't entirely save him from censure.) His contemporary, the Frenchman Julien Offray de La Mettrie, had no qualms in making us all mere mechanism, a position he maintained in his 1747 book *L'Homme machine*, which the church condemned as fit for burning.

You needn't be an anthropocentric bigot to take the view that mind was an abrupt evolutionary innovation. Maybe that leap happened for the common ancestors we share with great apes? Perhaps mind appeared with the origin of all mammals?

The proposition, however, seems unlikely. Evolutionary jumps and innovations do happen – but as we'll see, there's no sign in the evolutionary record of a transition to mindedness suddenly transforming the nature or behaviour of pre-human creatures. Nor is there any reason to think that the explosion, around forty to fifty thousand years ago, in the capabilities and complexities of the behaviour of *Homo sapiens* was due to the abrupt acquisition of mind itself. It looks much more probable that the quality that I propose to associate with mind arose by degrees over a vast span of evolutionary time. There's now a widespread view that it was present

to some extent before our very distant ancestors had even left the sea. If so, there is perhaps nothing any more special about it than there is about having a backbone, or breathing air.

In either scenario, it's by no means obvious that mindedness need confer an adaptive benefit at all. Could it be that this attribute, which strikes us as so central to our being (and surely it is, not least in being the quality that allows us to recognize it and be struck by it), was just a side effect of other cognitive adaptations? In other words, could it be that, if we and other creatures are to have brains that are able to do what they do, we have no option to incur a bit of mindedness too?

If that seems an alarming prospect – that evolution was indifferent (initially) to this remarkable and mysterious property that matter acquired – so too might be the corollary: perhaps matter could develop all kinds of capabilities for processing, navigating and altering its environment while possessing no mindedness at all. After all, a great deal (not all!) of what we find in the characteristics of life on Earth is highly contingent: the result of some accident or chance event in deep time that affected the course of all that followed on that particular branch of the tree of life. Could it be that evolution might have played out just as readily on Earth to populate it with a rich panoply of beings, some as versatile and intelligent as those we see today – yet without *minds*?

These could seem like idle speculations, fantastical might-have-beens that we can never go back and test. But by exploring the Space of Possible Minds, we can make them more than that.

What's the brain got to do with it?

Neuroscience barely existed as a discipline when Gilbert Ryle wrote *The Concept of Mind*, but he doubted that the 'science of mind' then

prevailing – psychology* – could tell us much beyond rather narrow constraints. It was no different, he said, from other sciences that attempt to categorize and quantify human behaviour: anthropology, sociology, criminology, and the like. There is no segregated field of mental behaviour that is the psychologist's preserve, he said, nor could it ever offer causal explanations for all our actions in the manner of a Newtonian science of mind. At root, Ryle's scepticism towards a 'hard-science' approach derives from the central problem for understanding the mind: we can only ever come at it from the inside, which makes it different from studying every other object in the universe. That's one way of expressing what is often called the 'hard problem' of consciousness: why a mind has anything there is to be like. We can formulate testable theories of why the brain might generate subjective experience, but we don't know even how to formulate the question of why a given experience is like *this* and not some other way: why red looks red, why apples smell (to us) like apples. (It might not, as we'll see, even be a question at all.)

Ryle is surely right to suggest that some problems of mind are irreducibly philosophical. But he threw out too much. To recognize that there are limits to what the brain and behavioural sciences can tell us about the mind is not the same as suggesting that they can tell us nothing of value. Indeed, to talk about minds without consideration of the physical systems in which they arise is absurd, akin to the sort of mysticism of mind that Ryle wanted to dispel.

So we need to bring the brain into the picture – but with care. The human brain is surely the orchestrating organ of the human mind, but the two concepts are not synonymous – for the obvious reason that the human mind didn't evolve solely for the sake of the brain, or vice versa. Minds as we currently know them belong to

* The discipline of course still exists, but now it overlaps considerably with what is commonly called cognitive science, and indeed with neuroscience.

living entities, to organisms as a whole, even if they are not distributed evenly throughout them like some sort of animating fluid.

I fear Ryle wouldn't like this perspective either. He derided the Cartesian division of mind from body as the 'ghost in the machine', and he argued instead that mind shouldn't be regarded as some immaterial homunculus that directs our actions, but is synonymous with what we do, and thus inseparable from the body. Yet he felt the problem was not so much that the two are intimately linked as that Descartes' dualism is a category error: minds are fundamentally different sorts of things from bodies. It is no more meaningful, he wrote, to say that we are made up of a body plus a mind than that there is a thing made up of 'apples plus November'. Descartes only bracketed the two together (Ryle says) because he felt duty-bound, in that age, to give an account of mind that was couched in the language of mechanical philosophy: the body was a kind of mechanism, and so the mind had to be something of that kind too, or related to it. Ryle would probably take the same dim view of modern neuroscience, which seeks to develop a mechanistic account of the human brain. Yet the simple fact is that no one can (or at least, no one should) write a book today about the question of minds while excluding any consideration of neuroscience, brain anatomy and cognitive science. Nor, for that matter, can they ignore our evolved nature. It is like trying to talk about the solar system without mentioning planets or gravity.

The brain, though, is a profound puzzle as a physical and biological entity. Compare it, say, with the eye. That organ is a gloriously wrought device,* including lenses for focusing light, a moveable

* Wrought, I hope it goes without saying, by the blind forces of evolution, which sift random changes in form for ones that improve function, in ways advantageous to survival and the propagation of offspring. The marvellousness of the eye has made it a favourite example for those who wish us to believe that it must be

aperture, photosensitive tissues to record images, delicate colour discrimination, and more. All of these components fit together in ways that make use of the physical laws of optics, and those laws help us to understand its workings. The same might be said of the ear, with its membraneous resonator and the tiny and exquisitely shaped bones that convey sound along to the coiled cochlea, capable of discriminating pitch over many orders of magnitude in frequency and amplitude. Physics tells us how it all functions.

But the brain? It makes no sense at all. To the eye it is a barely differentiated mass of cauliflower tissue with no moving parts and the consistency of blancmange, and yet out of it has come *Don Quixote* and *Parsifal*, the theory of general relativity and *The X Factor*, tax returns and genocide.

Of course, under the microscope we see more: the root network of entangled dendrites and their synaptic junctions, the mosaic of neurons and other cells, bundles of fibres and organized layers of nerves, bursts of electrical activity and spurts of neurotransmitters and hormones. But that in itself is of little help in understanding how the brain works: there's nothing here suggestive of a physics of thought in the same way as there is of vision and hearing. Conceivably, the microscopic detail just makes matters worse (at least at first blush), because it tells us that the brain, with its 86 billion neurons and 1,000 trillion connections, is the most complex object we know of, yet its logic is not one for which other phenomena prepare us.

What's more, the lovely contrivances of the ear and eye (and other facilitators of the senses) are in thrall to this fleshy cogitator. Though we can understand the physical principles that govern sight and sound, the brain can override them. It makes us see things that are patently absent, such as the light falling on the retina, and also

---

truly miraculous: that a divine intelligence lies behind this and other of nature's designs. But any need for such foresight has long since been proved otiose.

remain blind to things that imprint themselves there loud and clear. The output of the ear is like an oscilloscope trace of complex sonic waveforms: none of it comes labelled as 'oboe' or 'important command' or 'serious danger alert', and certainly none instructs us to feel sad or elated. That's the brain's job.

All this means that science can be forgiven for not understanding the brain, and deserves considerable praise for the fact that it is not still a total mystery. The best starting point is an honest one, such as can be found in Matthew Cobb's magisterial 2019 survey *The Idea of the Brain*, which states very plainly that 'we have no clear comprehension about how billions, or millions, or thousands, or even tens of neurons work together to produce the brain's activity.'

What we *do* know a lot about is the brain's anatomy. Like all tissues of the body, it is made up of cells. But many of the brain's cells are rather special: they are nerve cells – *neurons* – that can influence one another via electrical signals. It's easy to overstate that specialness, for many other types of cell also support electrical potentials – differences in the amount of electrical charge, carried by ions, on each side of their membranes – and can use them to signal to one another. What's more, neurons are like other cells in conveying signals to one another via molecules that are released from one cell and stick to the surface of another, triggering some internal change of chemical state. But only neurons seem specially adapted to make electrical signalling their *raison d'être*. They can achieve it over long distances and between many other cells by virtue of their shape: tree-like, with a central cell body sporting branches called dendrites that reach out to touch other cells, and an extended 'trunk' called an axon along which the electrical pulse (a so-called action potential) can travel (Figure 1.1). Each of these pulses lasts about a millisecond.

This 'touching' of neurons happens at junctions called synapses (Figure 1.2), and it doesn't require physical contact. Rather, there is

Figure 1.1. The structure of neurons and the synaptic junctions between them.

a narrow gap between the tip of an axon and the surface of another neuron's dendrite with which it communicates, called the synaptic cleft. When an electrical signal from the axon reaches the synapse, the neuron releases small biomolecules called neurotransmitters, which diffuse across the synaptic cleft and stick to protein molecules called receptors on the surface of another cell – these have clefts or cavities into which a particular neurotransmitter fits a little like a key into a lock. When that happens, the other cell's electrical state changes. Some neurotransmitters make the cell excited and liable to discharge a pulse of their own. Others quieten the cells they reach, suppressing the 'firing' of the neuron's distinctive electrical spike.*

* Different neurotransmitters are often described as though they have specific effects on the brain: serotonin is the 'well-being' signaller, suppressing violence

Figure 1.2. Close-up in the synaptic junction. Communication from one neuron to another happens when neurotransmitter molecules are released from the tip of the axon into the narrow gap between it and the dendrite of another neuron. These molecules diffuse across and bind to special 'receptor' molecules on the surface of the dendrite.

In this way, the network of neurons becomes alive with electrical activity. Typically the brain develops synchronized patterns of neural firing, and this synchrony seems central to coherent cognitive

---

and aggression, dopamine is the 'euphoria' signaller, and so on. But while it's surely true that different neurotransmitter molecules have different roles and effects, we should resist stereotyping them, in much the same way that we should resist labelling certain genes as being responsible for specific traits. Human-level experience and behaviour rarely if ever translate in a transparent way down to the level of molecules.

processing. The activity is influenced by the signals the brain receives from sensory organs and nerves elsewhere in the body – those, say, from the optic nerve connected to the retina of the eye. In this way the raw information registered by such sensory organs is somehow turned into mental images: thoughts, feelings, memories.

It's often overlooked that the brain's activity doesn't *rely* on such stimuli. It is intrinsically active: neurons are communicating with each other all the time. We don't know what they are 'saying', but the activity is not random, and it seems likely to be a vital part of cognition. In fact, the signals caused by sensory stimuli are typically very small by comparison, so that they can be detected by brain-monitoring technologies only after averaging away all the brain's noisy (but not merely random) 'background chatter'.

There are hundreds of different types (and many more sub-types) of neurons, each differing in size, shape and patterns of electrical activity. Some are excitatory: they stimulate others to fire – while others are inhibitory, suppressing activity in those to which they are connected. And brains are not just neurons. There are other cell types present too, in particular so-called glial cells, which outnumber neurons by a factor of about ten. Once considered just a kind of 'glue' tissue (that's what the name glia means) to bind the neurons together, glial cells are now known to play several vital roles in brain function, for example helping to maintain and repair neurons.

Too often overlooked also is the brain's energy economy. Our brains are big and expensive. Sheer physical size is not quite the issue – the sperm whale's brain is several times heavier (around 7.5kg) than ours, and men have a brain volume around 10 per cent larger on average than that of women without any difference in intelligence. What matters more is the ratio of brain mass to body mass, which is greater for humans than for any other animals. (Dolphins and great apes come next.) The brain accounts for about a

quarter of the energy consumption in an adult (for newborn babies it is around 87 per cent), and the energy demand is almost twice as great when the brain is conscious and aware than when it is rendered unconscious by anaesthesia: thinking literally demands brain power. One of the marvels of the brain is that it doesn't simply fry itself with all this energy use: a computer with this density of components and interconnections would simply melt from the heat it would have to dissipate.

Although the general anatomy of the brain is as pre-determined as any other part of the body by the interaction and activation of genes and cells during development, the details of the wiring are shaped by experience. This happens in a process more akin to sculpture than to painting. Rather than the neural network growing gradually link by link, it starts as a profuse abundance of branches and junctions that gets pruned back. At birth, a baby's brain typically contains more than 100 billion neurons, which is somewhat more than the adult brain. Connections that are unused wither away, while recurring patterns of activity are reinforced. Neurons that are activated by a particular stimulus tend to forge connections with one another, creating circuits with specific functions.

This process continues throughout growth and life in a constant feedback between brain and environment. It's one reason why we can't easily or meaningfully separate what is innate in a brain from what is acquired by experience, for pre-existing traits may become amplified. A child with inherently good pitch perception or rhythmic coordination is encouraged to study music and thereby improves these traits still further, strengthening the relevant neural circuits. Any small gender-related differences in cognition – and what these are, or whether they exist at all, remains controversial* – are likely to

* Humans are in fact unusual among animals in having so few clear-cut gender differences in behavioural traits and abilities. Neuroscientist Kevin Mitchell

be amplified by cultural stereotypes, even in the most enlightened households. What's more, the traits we might recognize at a behavioural level, such as neuroticism or openness to new experiences, don't seem to have any clear correlates at the level of the neural circuitry – there aren't, say, 'conscientiousness' neurons. Personality is evidently a meaningful notion at the social level, but it's not obviously to be found in the wiring of the brain. Rather, these high-level, salient aspects of behaviour arise out of more 'primitive' characteristics, some of which *do* seem to have their origin in specific, identifiable aspects of the brain such as hormone levels or production of different neurotransmitters: aversion to harm, say, or ability to defer gratification.

Other traits and abilities are, however, more fundamental. The basic abilities that enable us to make sense of and navigate the world, such as vision, language, and motor skills, are each produced in specific regions of the brain. But not necessarily unique ones. Vision, for example, begins when light on the retina of the eye stimulates the optic nerve to send signals to the primary visual cortex (located, oddly, at the back of the brain). But as we'll see, for us to perceive an object visually it's not enough that its visual signal arrives here. The primary visual cortex communicates with other regions of the cortex that *interpret* what is seen, for example by separating it into different components (such as edges, textures, shadows, and movement) and interpreting them in terms of familiar objects. Vision defects can result from malfunctions at any of these stages.

---

speculates that perhaps some anatomical differences (on average) between male and female brains exist to *compensate* for the very clear physiological differences, for example in brain size or hormones, that might otherwise be expected to produce cognitive differences. Maybe there were good evolutionary reasons for human males and females to become more alike in their behaviour.

The brain has a mystique unequalled by any other part of the body. We commonly imagine that it holds the secrets of all that we are. There is marginally more justification for that notion than there is for the other modern biological myth of identity, which ascribes it all to the genome. For while our genes can dictate various dispositions in how the brain (as well as the rest of the body) develops, unquestionably influencing our traits, innate abilities and behaviours, the mantra that 'DNA is not destiny' can't be stressed enough.

It's a common view that personality and character traits arise through a combination of nature – genetic predispositions – and nurture, the slings and arrows of experience. Both play a role, but they are hard to disentangle and even harder to predict. Discomfiting though many find that idea, the clear fact is that genes play a significant, perhaps sometimes even dominant role, in guiding what we do. There is no human trait, from sexual orientation to a propensity for watching television or probability of divorcing, that doesn't appear to have a genetic component to it.* Yet environment or experience too may affect the way the brain gets wired up. Training for a skill, for example, can restructure the parts of the brain concerned, while deprivation or abusive treatment can leave long-lasting scars on the psyche – which necessarily means on the physical structure of the brain. Many environmental effects are hard to predict, however. Particular life events – a parental divorce, say – can have very different effects on different individuals, in part

* It sounds unlikely, even absurd, that there should be any genetic basis for television-watching, given that television is much too recent an invention for natural selection to have exerted any influence on watching habits. But as with so many other aspects of behaviour, the determining factors here are very general ones on which selection has long acted, such as an ability to maintain focused attention.

because of the innate characteristics of personality. Bluntly put, some people will weather trauma better than others.

There is a third influence, however, on the individual features of the brain (and the mind it supports) that is too often overlooked: neither nature (genes) nor nurture (experience), but chance events in both development and the environment. Genes give the relevant cells a programme of sorts for how to organize themselves into a brain structure, and they can bias certain aspects of that organization. But they cannot fully prescribe it, not least because the number of neural connections outweighs the number of genes involved in brain development by several billionfold. There is no blueprint for a brain in the genome. (In truth there's no blueprint for anything.)

The developmental process of growing a brain is thus astronomically complex, and tiny, random events during its course can have significant knock-on effects. If we have a particular talent, we love to say that it's in the genes, and seek for a relative to whom we can give the credit. Perhaps that's so; but it's often also possible that you just got lucky in the way your brain happened to wire up.

Is the brain a mind machine?

The brain's mystique surely stems from the notion that from it comes all that we are. The body, in this view, is just the housing: the mindless machinery that the brain controls. Thence comes the fascination with famous brains – as for example in the determination of some surgeons to seek the origin of Einstein's genius in slices of his grey matter, preserved against his wishes after his death, some of which circulated within medical networks like contraband.

There is a lot invested in the human brain. It is reasonably considered the 'engine of mind', provided that we remain alert to the shortcomings of such a mechanical metaphor. It both determines

and constrains what a human mind can be like. But it does not embody the sum total of that issue, nor can it answer all the questions we might ask about mindedness. In exploring the Space of Possible Minds, we will need to perform a constant dance between the physical 'hardware' and the properties that emerge from it – a dialogue of mind and matter.

When we reason, think, feel, there are of course brain circuits underpinning those activities, and occasionally I will need to refer to them. But it's important to keep in mind that how the mind works and how the brain works are two distinct (though related) questions. Understanding one of them doesn't necessarily tell us about the other, and much of neuroscientific research today is aimed at trying to connect the two.

Neuroscience now has a wonderful battery of techniques for studying brain activity, most notably functional magnetic resonance imaging (fMRI), which reveals where blood flow has increased because of the demands for oxygen created by cell (neural) activity. These maps of brain activity can be revealing. If for example we discover that a specific cognitive task involves a part of the brain previously associated with a different function – if, say, listening to music activates regions linked to language processing – that can give us insight into the kind of mindfulness that is going on: the brain may be identifying a sort of syntax or grammar in the music. On the other hand, very often such regions have been labelled in the first place because of their activation in response to a particular task or behaviour, rather than because we know what's happening in said region. Unfortunately, brain-imaging methods don't come supplied with a convenient description of what the associated neural circuits are actually doing with the information they receive.

The danger is that the brain becomes simplistically divvied up into regions ascribed to specific functions, in much the same way as it was in the now discredited late nineteenth-century 'science' of

phrenology. That tendency has given rise to simplistic and misleading tropes about brain anatomy: that the structure called the amygdala is said to be the 'fear centre', say. Such habits are encouraged by some neuroscientists' unfortunate tendency to confuse an explanation of a cognitive task with a mere list of the obscurely named parts of the brain that are active during it: the anterior cingulate cortex, the parahippocampal gyrus, and so on.

This returns us to the mind–brain problem: we can examine the brain as a biological organ all we like, but we still can't get inside the mind it helps create, and see what is going on. This is a tricky notion to grasp. Without the human brain, there is no human mind – we can be confident about that.* But does this mean that the brain is all there is to mind, or could it be analogous to saying that without the conductor, the orchestra does not play? The conductor is the organizer, the hub of all the activity – but heck, the conductor does not even play the music. Usually the conductor did not even compose the music. In fact, where *is* the music – on paper, in the vibrations of the air, or in the minds of the audience? What generates the music of the mind? And how many kinds of music can it play?

As we have seen, the brain–mind relationship is commonly couched today in terms of an analogy with computers: the brain is said to be the hardware (or perhaps, the 'wetware') and the mind is the software. In this picture – widely, if not universally, held in neuroscience – the brain is itself a kind of computer.

The analogy has existed since the earliest days of computers: the pioneer of information technology John von Neumann wrote a book titled *The Computer and the Brain* in 1957. Brains, as we'll see, have often been the explicit inspiration for work in computer

* Whether all minds require brains is another (and unresolved) matter, as we'll see.

science and artificial intelligence, to the point where, according to neuroscientist Alan Jasanoff, 'it can be difficult to tell which is the inspiration for which.' The computational view of mind is agnostic about where we might expect to find minds: it allows mindedness to arise in any system capable of the requisite computation. It represents one view of the philosophical perspective on mind called functionalism, which holds that a mental state or operation – a thought, desire, memory – doesn't depend on what the mind is made from, but on what it does – on its functional role in the mind-system, independent of the substrate in which it is supported. That view can be discerned in Thomas Hobbes' mechanistic view of the human mind, whereby it is all a kind of arithmetical processing: reason, he wrote in 1651, is 'nothing but *reckoning*, that is adding and subtracting.'

You can see the attraction of this idea. Just as computer circuits process information by passing digital (on/off) electrical signals through networks of vast numbers of interlinked 'logic gates' made from transistors, so the brain has neural networks made from interlinked neurons whose action potentials switch one another on and off. The brain, the argument goes, is much more powerful and versatile than the computer because it has many more components and connections; but the computer is often faster, because silicon devices can switch in nanoseconds (billionths of a second) while neurons fire only a few times per second.

One of the reasons the brain–computer analogy is so popular is that it offers what may be a false reassurance that the brain can be understood according to well-established engineering principles. It might be seen, Jasanoff suggests, as the modern equivalent of Descartes' separation of messy, material body from pristine, immaterial mind. The corollary is that mind doesn't need brain at all, but just information: a computer can simulate a brain in every detail, and thereby host a mind. 'Equating the organic mind to an inorganic

mechanism might offer hope of a secular immortality', says Jasanoff. I'll return to this fantasy later.

But does the analogy hold up? History alone should make us wary: metaphors for the brain have always drawn on the most advanced technologies of the day. Once they were regarded as working like clockwork, or hydraulics, or electrical batteries. All of these now look archaic to us, and it's possible that the computer comparison will too, one day.

The most fundamental objection to the analogy is that brains and minds simply do not do what computers do. Gilbert Ryle considered it a common mistake to 'suppose that the primary exercise of minds consists in finding the answers to questions'. That's typically the task for a computer: it must compute an output based on the values input to a well-defined algorithm. It *looks* as though the brain does something like that too – that it enacts an algorithm for turning the input data of sensory experience into outputs such as actions. But the resemblance is superficial. Rather than information flowing through the brain in one direction from input to output, artificial-intelligence expert Murray Shanahan observes that 'waves of activation' can move back and forth between different regions until settling into a 'temporary state of mutual equilibrium.' The brain does not sit there waiting for questions to answer, but is in constant, active and structured conversation with itself and its environment. Frankly, we don't yet understand much about that discourse.

Moreover, the computer analogy says nothing about the role of sentience, which is surely one of the most salient aspects of our own minds. We can already make machines – assuredly non-sentient entities, at least at this stage of their development – that function as information-processing input–output devices that can decide on and execute actions. Conversely, our own minds seem capable of experiencing without acting: of being in states close to pure awareness, without any objectives or outputs to compute.

What's more, we can't in general, enumerate the reasons why we do what we do by mapping out the logical processes of mind in the way we can for a computer's circuits – identifying the input data, the key decision points and the outcomes of those processes, and so on. (We *do*, however, devote a lot of energy to constructing narratives of that sort – because coherent stories are something our minds appear to crave, for good adaptive reasons).

Yes, the mind has to make decisions conditioned on existing and new information – but those decisions are not what *make it a mind*. They are not what gives to a mind something that it is to be like. As we'll see, it looks possible that the subjectivity of awareness might be fundamental to what a mind does, not just an epiphenomenon of its problem-solving nature. Neither is the mind a kind of machine that goes to work on its environment; rather, it is something that emerges from the interaction of organism and environment. For this reason, it is neither some ethereal or abstract essence, nor a thing that can be written down as a kind of code or algorithm. It is a process, perpetually running, forever recreating and updating itself. The brain of a moving, active organism is part of a constant feedback loop: it controls and selects which stimuli it perceives, as well as adjusting to those stimuli. Typically, experiments on behaviour open up that loop artificially by selecting stimuli for the brain and watching the result. But ordinary behaviour is quite different, which is why cognition – what the brain actually does in the environment, an active and ongoing process – is not passive input–output computation. The 'input–output doctrine', writes neuroscientist Martin Heisenberg, 'is the wrong dogma, the red herring [in brain research].'

Even reflex actions – where stimulation of a given nerve or set of nerves automatically generates a particular response, like the patellar reflex where a tap of the knee produces an automatic leg jerk – are in fact controlled by the brain and not just passively enacted by

them. What might look like a simple input–output response may actually be fine-tuned by the brain in response to the prevailing circumstances. Thus, some apparently instinctive and unmeditated responses in animals are actively altered and directed by the brain in subtle ways to suit the situation: the organisms are not acting with the predictability of a light switch.

What's more, brains do not need stimuli in order to be active. They are busy during sleep, for example, when our senses are mostly shut down. While external realities (coldness, or a full bladder) might sometimes leak into our dreams, most of their content is internally generated. Turn off all the stimuli to an awake and conscious person – place them in a sensory-deprivation chamber – and the mind will soon start to produce its own activity, which manifests as hallucinations. Nervous systems are dynamic and constantly active, whether or not the world feeds them sensory data. Minds exist *to mind*.

In contrast with a conventional computer algorithm, then, you cannot work out what a human brain will do when given a certain set of sensory inputs (which would have to include hormones and other bodily signals).* After all, the possible configurations and states that can be sustained in a typical neural network of a brain is larger than the number of particles in the observable universe. Yet just a tiny fraction of these states have been selected by evolution, and they are reliably produced in response to pretty much whatever stimuli the brain receives.†

* We'll see later that the approach generally used for today's artificial intelligence, called machine learning, also lacks transparent predictability in how inputs produce outputs. But we'll also see why this computational process differs from the way brains work.

† Are there any stimuli that the brain simply can't handle – for which it can't settle into one of its selected states? It would be intriguing, though perhaps hazardous,

That's the key reason why neuroscientists Gerald Edelman and Giulio Tononi argue that the brain can't be considered a computer. For even though many aspects of cognition and neural processing do closely resemble computation, it does not arrive at its internal states by logic operations, but by *selection*.

This might not be a coincidence. That's to say, it might be that the minds of evolved living beings *have* to be of this type, because purely logic-based machines won't work. Living organisms need, for one thing, to have some stability of behaviour – a reliable repertoire of possible actions adequate for most situations they will face, even though those are unpredictable and never identical. And they need to make decisions quickly, without laborious computation of all the input signals. But perhaps most of all: they are decision-making devices that *don't really have a well-defined input at all*. There's no moment (except perhaps in maths tests) where our mind is presented with a closed problem, characterized by uniquely specified, single-valued parameters, that we have to solve. The world is not a computer tape that feeds a binary digit string into our brain. 'Like evolution itself', say Edelman and Tononi, 'the workings of the brain [it might be better to say 'of the mind'] are a play between constancy and variation, selection and diversity.' Machines don't work this way – or perhaps it is better to say, we've never yet made machines that do. Computer scientist Josh Bongard and biologist Michael Levin turn the matter on its head, suggesting that perhaps living and minded entities show us 'machines as they could be'.

Yet my hunch is that no genuine mind will be like today's computers, which sit there inactive until fed some numbers to crunch.

---

to know. Might we indeed have a mental meltdown when faced with the cosmic infinitude of H. P. Lovecraft's Chthulu, or when fed the wisdom of super-advanced intelligences, as two hapless characters are in Fred Hoyle's sci-fi novel *The Black Cloud* (page 338)?

It's certainly notable that evolution has never produced a mind that works like such a computer. That's not to say it necessarily cannot – evolution is full of commitment bias to the solutions it has already found, so we have little notion of the full gamut it can generate – but my guess is that such a thing would not prove very robust. Edelman and Tononi postulate that logical computation (as in our artificial 'thinking devices') and selection (in evolved minds) might be the *only* two 'deeply fundamental ways of patterning thought'. If so, that would surely be a fundamental aspect of topography of the Space of Possible Minds. But *are* they the only two? We know of no others, but personally I'd be surprised if they don't exist. After all, we are just starting our journey into Mindspace, and we have very little conception of how big it is, and what lies out there.

CHAPTER 2

The Space of Possible Minds

In 1984 the computer scientist Aaron Sloman, of the University of Birmingham in England, published a paper arguing for more systematic thinking on the vague yet intuitive notion of mind. It was time, he said, to admit into the conversation what we had learned about animal cognition, as well as what research on artificial intelligence and computer systems was telling us. Sloman's paper was titled 'The structure of the space of possible minds'.

'Clearly there is not just one sort of mind', he wrote:

> Besides obvious individual differences between adults there are differences between adults, children of various ages and infants. There are cross-cultural differences. There are also differences between humans, chimpanzees, dogs, mice and other animals. And there are differences between all those and machines. Machines too are not all alike, even when made on the same production line, for identical computers can have very different characteristics if fed different programs.

Now an emeritus professor, Sloman is the kind of academic who can't be pigeon-holed. His ideas ricochet from philosophy to

information theory to behavioural science, along a trajectory that is apt to leave fellow-travellers dizzy. Ask him a question and you're likely to find yourself carried far from the point of departure. He can sound dismissive of, even despairing about, other efforts to ponder the mysteries of mind. 'Many facts are ignored or not noticed,' he told me, 'either because the researchers don't grasp the concepts needed to describe them, or because the kinds of research required to investigate them are not taught in schools and universities.'

But Sloman shows deep humility about his own attempt four decades ago to broaden the discourse on mind. He thought that his 1984 paper barely scratched the surface of the problem and had made little impact. 'My impression is that my thinking about these matters has largely been ignored', he says – and understandably so, 'because making real progress is very difficult, time-consuming, and too risky to attempt in the current climate of constant assessment by citation counts, funding, and novel demonstrations.'

But he's wrong about that. Several researchers at the forefront of artificial intelligence now suggest that Sloman's paper had a catalytic effect. Its blend of computer science and behaviourism must have seemed eccentric in the 1980s but today it looks astonishingly prescient.

'We must abandon the idea that there is one major boundary between things with and without minds', he wrote. 'Instead, informed by the variety of types of computational mechanisms already explored, we must acknowledge that there are *many* discontinuities, or divisions within the space of possible systems: the space is not a continuum, nor is it a dichotomy.'

Part of this task of mapping out the space of possible minds, Sloman said, was to survey and classify the kinds of things different sorts of minds can do:

> This is a classification of different sorts of abilities, capacities or behavioural dispositions – remembering that some of the behaviour may be internal, for instance recognizing a face, solving a problem, appreciating a poem. Different sorts of minds can then be described in terms of what they can and can't do.

The task is to explain what it is that enables different minds to acquire their distinct abilities.

'These explorations can be expected to reveal a very richly structured space', Sloman wrote, 'not one-dimensional, like a spectrum, not any kind of continuum. There will be not two but many extremes.' These might range from mechanisms so simple – like thermostats or speed controllers on engines – that we would not conventionally liken them to minds at all, to the kinds of advanced, responsive, and adaptive behaviour exemplified by simple organisms such as bacteria and amoebae. 'Instead of fruitless attempts to divide the world into things with and things without the essence of mind, or consciousness', he wrote, 'we should examine the many detailed similarities and differences between systems.'

This was a project for (among others) anthropologists and cognitive scientists, ethologists and computer scientists, philosophers, and neuroscientists. Sloman felt that AI researchers should focus less on the question of how close artificial cognition might be brought to that of humans, and more on learning about how cognition evolved and how it manifests in other animals: squirrels, weaver birds, corvids, elephants, orangutans, cetaceans, spiders, and so on. 'Current AI', he said, 'throws increasing memory and speed and increasing amounts of training data at the problem, which allows progress to be reported with little understanding or replication of natural intelligence.' In his view, that isn't the right way to go about it.

What it is like

Although Sloman's concept of a Space of Possible Minds was stimulating to some researchers thinking about intelligence and how it might be created, the cartography has still scarcely begun. The relevant disciplines he listed were too distant from one another in the 1980s to make much common cause, and in any case we were then only just beginning to make progress in unravelling the cognitive complexities of our own minds. In the mid-1980s, a burst of corporate interest in so-called expert-system AI research was soon to dissipate, creating a lull that lasted through the early 1990s. The notion of 'machine minds' became widely regarded as hyperbole.

Now the wheel has turned, and there has never been a better time to consider what Sloman's 'Mindspace' might look like. Not only has AI at last started to prove its value, but there is a widespread perception that making further improvements – and perhaps even creating the kind of 'artificial general intelligence', with human-like capabilities, that the field's founders envisaged – will require a close consideration of how today's putative machine minds differ from our own. Understanding of animal cognition too has boomed in the past two decades, in part because of the new possibilities that neuroscience and information technologies have opened up (but frankly, mostly because of better behavioural experiments). Child-psychologists now routinely talk to roboticists and computer engineers, and neurologists to marine biologists. We have some of the conceptual and experimental tools to start mapping a landscape of minds.

To avoid raising false expectations, however, I must confess at once that neither I nor anyone else yet knows what the natural coordinates of Mindspace are – nor even whether they are well-defined at all. Minds seem unlikely to be the kinds of objects that

one can represent by mathematical functions of precisely quantifiable variables: Mind = $xy^2 + qz^3$ or some such. Yet I do believe we can identify some of the likely *components of mindedness*: the qualities that minds seem to exhibit. What's more, we can anticipate that some minds are likely to be more richly imbued with these qualities than others. We can ask what kinds of minds, as a consequence, they are. Perhaps we can even try to imagine our way inside them.

To consider this Mindspace most fruitfully, I believe we will need to exercise our imagination. That facility is a vastly underrated tool in science, which (from the outside at least) can look like a pursuit built from strict logic and rigour, demanding precision of theory and observation. From the inside, on the other hand, I suspect there is no scientist who does not appreciate the value of imagination. Many like to cite Einstein's famous quote (which for once is genuine):

> I'm enough of an artist to draw freely on my imagination, which I think is more important than knowledge. Knowledge is limited. Imagination encircles the world.

What Einstein was referring to here is something like the leap of faith that enables the scientist to see just beyond the data, to construct new hypotheses for empirical testing, and to trust in the intuition that might be required to sustain their speculative ideas while no data exist to adjudicate on them.

The imagination that will help us to explore the Space of Possible Minds is of a somewhat different order, more akin to that required by a novelist: the ability to see through other eyes, to hear with other ears. With skill and research, it is possible to make a good fist of imagining oneself into the head of a Renaissance painter in Florence or a courtier of the Chinese Tang dynasty – or (if you are

not already one yourself) into the viewpoint of an autistic boy, as in Mark Haddon's 2003 novel *The Curious Incident of the Dog in the Night-time*. On occasion, novelists have tried to go further: James Joyce projecting himself into early childhood at the start of *Portrait of an Artist as a Young Man*, Laline Paull's apian protagonist in *The Bees*, Orhan Pamuk making his narrator a coin or the colour red in *My Name is Red*. Yet there's little attempt in these latter works to make the characters anything more than humanized non-humans.

No dog, let alone a coin, can write a book, and no ghost-writer can help with that. One of the most gloriously quixotic attempts to install oneself in a non-human mind is Charles Foster's *Being a Beast* (2016), in which he tries to enter into the lifestyle and perspective of an otter, a fox, and a swift. In the end, Foster's book reveals how doomed an enterprise that is. But if there's one attribute that distinguishes the human mind from those of other creatures, it is our imagination: our capacity for fantastical 'what ifs', for mental metempsychosis. It's almost as if we have evolved to make these attempts to become what we are not. Of course we will fail if, say, we try to picture what it must be like to 'see' the world as a silicon-based neural network does, or a fruit fly. But what I hope to show you is that we can fail better – and to do so, we don't need to start from scratch, nor to set out totally blind into the landscape where our destination lies.

Anyone even vaguely familiar with the philosophy of mind will doubtless be wondering when Thomas Nagel's bat will flap into view. This seems to be its cue. In 1974 Nagel, a distinguished American philosopher, published what might be the most widely cited paper in the entire field of the philosophy of mind, titled 'What is it like to be a bat?'.

As you can see, the title reflects the very definition I have chosen for an entity to have mindedness: there is something it is like to be

that entity. Nagel chose to pose the question for a bat partly because this creature is evolutionarily distant enough from us to seem rather alien, yet close enough for most people to believe that it does have some subjective experience. (We will look at whether that sort of belief is justified, and what it might mean, in Chapter 5.) But he also selected the bat because the modality of its sensory environment is so different from ours: although it has a visual system, it creates an image of its surroundings primarily using sonar echolocation, the reflection from surfaces of the acoustic waves it emits as shrill shrieks.

Bat sonar, Nagel wrote,

> though clearly a form of perception, is not similar in its operation to any sense that we possess, and there is no reason to suppose that it is subjectively like anything we can experience or imagine. This appears to create difficulties for the notion of what it is like to be a bat. We must consider whether any method will permit us to extrapolate to the inner life of the bat from our own case, and if not, what alternative methods there may be for understanding the notion.

He went on:

> It will not help to try to imagine that one has webbing on one's arms, which enables one to fly around at dusk and dawn catching insects in one's mouth; that one has very poor vision, and perceives the surrounding world by a system of reflected high-frequency sound signals; and that one spends the day hanging upside down by one's feet in an attic. *In so far as I can imagine this (which is not very far)*, it tells me only what it would be like for me to behave as a bat behaves. But that is not the question. [My italics.]

It is, rather, the technique of the novelist, the modus operandi of, say, Richard Adams for imagining his rabbit protagonists in *Watership Down*. But, wrote Nagel,

> I want to know what it is like for a bat to be a bat. Yet if I try to imagine this, I am restricted to the resources of my own mind, and those resources are inadequate to the task. I cannot perform it either by imagining additions to my present experience, or by imagining segments gradually subtracted from it, or by imagining some combination of additions, subtractions, and modifications.

His conclusion that we cannot possibly hope – by our very nature, indeed essentially by our very definition – to know 'what it is like to be a bat' is hard to refute, and widely accepted. There's less consensus about the corollary Nagel presents, which is that consciousness itself – as he put it, the problem of how the mind arises from the body – is not accessible to scientific study:

> For if the facts of experience – facts about what it is like for the experiencing organism – are accessible only from one point of view, then it is a mystery how the true character of experiences could be revealed in the physical operation of that organism. The latter is a domain of objective facts par excellence – the kind that can be observed and understood from many points of view and by individuals with differing perceptual systems. There are no comparable imaginative obstacles to the acquisition of knowledge about bat neurophysiology by human scientists.

Even if we know all there is to know about the bat brain and physiology, that information can never tell us what it is to be a bat. By the same token, Nagel wrote that 'Martians might learn more about the human brain than we ever will' (he did not, I assume,

actually believe there are Martians; they play the role of generic super-smart aliens) yet never know what human experience is like.

I don't want to give the impression that the only way to explore Mindspace is to project ourselves imaginatively into it and take a stroll. As Nagel says, we'd be fooling ourselves to think we could ever achieve that in other than a superficial and schematic manner. My point is that, merely by being open to the notion that there are other things it is to *be like* cuts some of the tethers that the philosophy of mind creates when it supposes not just that the human mind is the most interesting case to focus on, but that it is the alpha and omega of the subject.

The mind club

Discussions about varieties of mind have tended to be either free-wheeling and amorphous or narrow and prescriptive. An example of the former is a 'taxonomy of minds' drawn up by Kevin Kelly, founding executive editor of the tech magazine *Wired*, which reads as an off-the-top-of-the-head list in which each item sounds like the premise of a sci-fi novel:

Cyborg, half-human half-machine mind
Super logic machine without emotion
Mind with operational access to its source code
Very slow 'invisible' mind over large physical distance
Nano mind

and so on.

Alternatively, minds might be classified according to some single measure, most obviously 'intelligence'. Frankly, we don't know what we mean by intelligence – or rather, we don't all, or always, mean the same thing by it. Of course, we do measure *human* minds on

an intelligence scale: the 'intelligence quotient' or IQ scale devised in 1912, allegedly the ratio of a person's mental to chronological age,* and which has been the subject of controversy ever since. Whatever their pros and cons for humans, IQ scales are certainly no good for other animals, or machines. If today's AI, or a dolphin, were to get a rotten score in an IQ test, we'd rightly suspect that it's not because they have 'low intelligence' (whatever that means) but because we're applying the wrong measure of intelligence – or more properly, of mind.

What's more, many tests of intelligence, like IQ scores, evaluate not cognition per se but performance. There's some value in doing that, but we know even from the way humans work that it can matter a great deal to the performance of a task how the question is asked, and in what circumstances. It is notoriously hard to find a fair test for all minds – trivially, I'd not fare at all well in an IQ test if it were written in Arabic. Performance alone might reveal rather little about the *kind* of mind being tested – about its modes of reasoning, of representing the environment, its intuitions and emotions (if any), its range of skills not probed by the test. A popular scheme for evaluating novel biologically inspired machine intelligence in the 2000s was called the Cognitive Decathlon, which tested skills such as recognition, discrimination, memory and motor control. It's too simplistic but nevertheless captures the spirit of the enterprise to say that the criterion of intelligence in this test was more or less whether

* IQ is generally considered a good proxy for the *g* factor, a measure of 'general intelligence' introduced around the same time by the English psychologist Charles Spearman. This was supposed to capture the observation that people (Spearman considered children) who perform well in one cognitive arena often do in others too. Like IQ, it has proved controversial – famously, the evolutionary biologist Stephen Jay Gould criticized it for reducing intelligence to a number, a 'single series of worthiness' that discriminated against disadvantaged groups.

the machine can play chess well. I don't think even Mensa would accept such a limited measure for humans, so why for machines?

While all those demonstrable skills matter a great deal for practical purposes, it seems equally important to have some sense of the kinds of cognitive realm these artificial 'minds' inhabit. In 1976, the psychologist Nicholas Humphrey argued that 'what is urgently needed is a laboratory test of "social skill"'. And what exactly is that? 'The essential feature of such a test', said Humphrey, 'would be that it places the object in a transactional situation where he [sic] can achieve a desired goal only by adopting his strategy to conditions which are continually changing as a consequence partly, but not wholly, of his own behaviour.' The principle is surely right, although such features of mind are likely also to be highly sensitive to context, and unlikely to be captured by a single measure.

What we really need to appreciate the varieties of mind is some way of plotting the dimensions that define the ways in which minds can work: more or less what Sloman's Space of Possible Minds posits. A simple yet ingenious way to begin that ambitious project was devised by American psychologists Daniel Wegner, Heather Gray and Kurt Gray (no relation) in 2007. They simply *asked* people about other minds: what attributes do we *think* they have? They canvassed the views of around 2,500 participants about the perceived mental capacities of humans, animals, and other entities such as robots, companies, and supernatural agents: ghosts and God. What, the volunteers were asked, do you imagine the minds of (say) chimpanzees and babies are like?

Surprisingly, the responses could be boiled down to a Space of Minds that had just two key attributes, which the researchers labelled *experience* and *agency*. Here 'experience' means not (as in colloquial usage) how *much* one has experienced, but rather the mind's innate capacity for an 'inner life': for feelings such as hunger, fear, pain, rage, pleasure, and joy. It really alludes to a sense that

there is something it feels like to have such a mind; in other words, my working definition of mind demands non-zero measures of experience. 'Agency', meanwhile, connotes the ability to *do* things and accomplish goals, and to exercise control in doing so: to possess memory, say, or to plan and communicate.

So this Space of Possible *Perceived* Minds – what we humans imagine as the dimensions of mind – can be displayed as an ordinary two-dimensional graph on which each entity is represented as a data point with a certain amount of 'experience' and 'agency' (Figure 2.1). Those entities that have non-zero quantities of either or both attributes, suggested Wegner and colleagues, are the ones that we humans admit to the Mind Club.

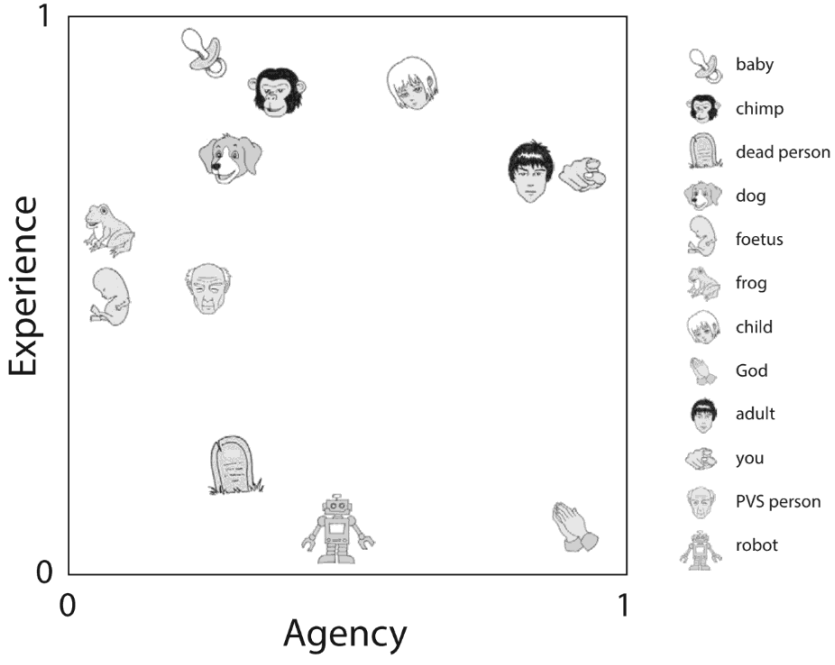

Figure 2.1. The Mind Club, according to Wegner, Gray and Gray: how we *think* about other minds.

It's something of a giveaway that adult humans are rated highly in both respects, and therefore appear at the top right of the graph. That's not to say we *don't* have a considerable amount of both experience (we feel, right?) and agency (we do stuff) – but it does suggest that all other minds are being assessed according to how they compare with us. That's hardly surprising. We have always thought in a human-centric way, and we shouldn't feel too bad about it. It is possibly in the nature, perhaps the very definition, of minds that they are at least somewhat egocentric and solipsistic. After all, when we started to map the cosmos we did that too from the human point of view: we put ourselves at the centre of all space. Literally, all else revolved around us – but what is more, it existed in relation to us and *for* us. It is us, not the beasts, that were deemed to be made in God's image and the primary focus of God's attention.

By degrees, our place in that cosmic expanse came to seem ever more contingent and insignificant. First we were on a planet like any other; then merely one of countless solar systems, embedded in countless galaxies – and perhaps even, some cosmologists think, inhabitants of one among countless universes. And just as the space of possible *worlds* looks very different today, so eventually will the Space of Possible Minds.

Humans are evidently regarded here as following a *path* in this space throughout the course of their lives. Babies were rated, on average, as slightly higher than adults on the ability to experience, but much lower on agency. Everything for a baby is intensely felt, eliciting happy gurgles or despondent howls; meanwhile, they can't get much done.* Children are intermediate between babies and

* Babies and even foetuses, like all complex organisms, have plenty of *biological* agency in the sense that I consider later (page 261) – but not in the more restrictive sense of the Mind Club.