4

Toward a Well-Innervated Philosophy of Mind

It might be thought that when I have been talking about 'philosophers of mind', collectively, as failing to appreciate our new folk neuroscience, in which the PNS is as important to conceptual issues as the brain (or more generally the CNS), I have neglected the newest developments in philosophy of mind and cognitive science, a group of proposals commonly referred to as *embodied and embedded* cognition (EE from now on). It is not a homogeneous group of theories, yet, all the proposals have in common a critique of cognitivism, that is, cognitive science that focuses on either abstract computational-symbolic phenomena in the brain, or on distributed, nonsymbolic, but still brain-confined patterns of nerve activation, and reduces or confines mental activity to these. The various EE proposals, starting from early 1990 (e.g., Varela, Thompson, and Rosch 1991),[1] have developed into a considerable literature by now, where the common point is to reestablish the mind-body-world connection that classical philosophy of mind and cognitive science have severed. There is not much space here to enter the details of EE, and though there is today a well-deserved sympathy for this approach, given its novelty and force, I will just point out that EE itself fails to

1990s

[1] Although some of the ideas emerged earlier, as pointed out by EE theorists, especially in the tradition of phenomenology established by Husserl and Merleau-Ponty, but also in the Heideggerian (or pop-Heideggerian, as the sarcastic critics sometimes refer to it) one.

meet the issues I have been discussing, namely the conceivability intuitions, head-on.

For instance, Andy Clark, in his latest book, *Supersizing the Mind* (2008), advertises Alva Noë's enactive or skill-based approach to experience (2004), based on what Noë calls 'sensorimotor knowledge' as "a powerful antidote to the venom of zombie thought experiments" (Clark 2008: 173). Yet, neither Clark nor Noë have *any* straightforward response to such thought experiments whatsoever. Instead, we find timorous references to the effect that as regards the 'phenomenal', it is, just maybe, not appropriate to equate it with the relevant naturalistic counterpart:

> All the vat scenario can directly establish is that, working together, the brain and the hyperintelligent vat conspire to support the usual panoply of cognitive and (*I am willing to venture*) phenomenal effects. (Clark 2008: 164) (emphasis added)

> A creature enjoys phenomenally conscious perceptual states when it has knowledge of the relevant patterns of dependence of neural activity on movement. But how can phenomenally unconscious states of this sort *be the basis of phenomenal consciousness*? This question remains unanswered. (Noë 2004: 228–29) (emphasis added)

The EE approach has all the elements needed to actually explain away, or at least weaken the zombie and other intuitions. For instance, the idea that it is an unsupported dogma to think of the brain as some kind of center of experience, and that experience is a global property of the whole nervous system, including the PNS, as it interacts with the world, could have been used, like in the case of the inconceivability of the zombie foot in the previous section, to show that the zombie intuition depends on some ways in which we tend to think of the subjects that instantiate experiences. Yet, those involved in the EE approach have failed to develop such arguments/thought experiments.

I will end with a few examples of problems in the philosophy of mind where changing from an exclusive focus on the brain to a closer attention to the PNS seems to bring about certain previously unexplored rejoinders.

I. 'It's Just Cables!'

Hilary Putnam's brain-in-a-vat thought experiment (1982) is one of the most discussed ones in current philosophy. Although it is mainly

used as an argument against skepticism, the scenario is also important when it comes to problems related to phenomenal consciousness. The scenario involves a brain that is artificially kept functioning in a vat (BIV from now on) and connected to a computer that stimulates it in such a way as to create an illusion of an external world of the kind we are experiencing: roads, mountains, lakes, forests, cars, our friends, and what not. However, all there really is around the brain is the vat and the computer that generates the signals. The brain will think thoughts; will believe that she/he is experiencing a real external world, interacting with the environment, and so on. Some philosophers appear to implicitly *define* phenomenal experience or phenomenology (not in the sense of the eponymous school of thought, but the way things appear to the subject of experiences) as 'whatever is shared between you and a BIV', that is, whatever is common to you as a subject of experience and the envatted physical copy of your brain as a subject of experience. Thus, Terry Horgan and John Tienson (2002), after reinforcing my earlier claim that philosophers think of the PNS as just cables and of experience as a terminus region in the brain, assume—as there is really *no* argument going on—that phenomenology is what you share with a BIV:

> Phenomenology does not depend constitutively on factors outside the brain. . . . First, phenomenology depends *causally* on factors in the ambient environment that figure as distal causes of one's ongoing sensory experience. But second, these distal environmental causes generate experiential effects only by generating *more immediate links in the causal chains between themselves and experience, viz., physical stimulations in the body's sensory receptors—in eyes, ears, tongue, surface of the body, and so forth*. And third, these states and processes causally generate experiential effects only by generating still *more immediate links in the causal chains between themselves and experience—viz., afferent neural impulses, resulting from transduction at the sites of the sensory receptors on the body*. Your mental intercourse with the world *is mediated by sensory and motor transducers at the periphery of your central nervous system*. Your conscious experience would be phenomenally just the same even if the transducer-external causes and effects of your brain's afferent and efferent neural activity were radically different from what they actually are—for instance, even if you were a Brain in a Vat with no body at all. (Horgan and Tienson 2002: 526–27) (emphasis added)

I have emphasized in the above quote the passages that indicate the bad folk neuroscience that we have been talking about in section 2 of chapter 3; but, besides that, what is going on when it comes to

phenomenology is an implicit definition of it via the notion of a BIV. A while before, John Searle (1983) had gone even further and had implicitly defined intentionality, or mental content, via the BIV, setting it as a condition that intentionality is whatever is shared between you and a BIV. Searle also believes in the brain as the seat of the person and experience; here is a famous quote in which he states his original view about the BIV:

> Each of our beliefs must be possible for a being who is a brain in a vat because each of us is precisely a brain in a vat; the vat is a skull and the 'messages' coming in are coming in by way of impacts on the nervous system. (Searle 1983: 230)

Several authors have expressed strong disagreement with Searle, and would, a fortiori, have expressed disagreement with Horgan and Tienson about our being in fact BIVs as far as phenomenology is concerned. Besides the traditional content-externalists, I'm thinking about the philosophers associated with the embodied cognition movement, and the extended mind hypothesis (Clark and Chalmers 1998). The idea is supposed to be that if the mind is constitutively embodied and embedded or situated in the environment, then an envatted, bodiless BIV will not necessarily have to share the mental life of our embodied and embedded minds. Recently, Clark offers reason to be skeptical about this whole line of thought (Clark 2009), and Chalmers (2005) puts forward the idea that the BIV hypothesis is a metaphysical rather than skeptical one, so that in the BIV scenario the underlying computational/informational patterns going on in the computer are to be equated with 'the world'. However, I want to draw attention to something else, which hasn't been noticed in the literature on BIVs.

I agree with Chalmers (2005) that it makes sense to think that the BIV has its own world, and it is no less reality than our world is; it (that world) just has a different fundamental informational/computational level that underlies it, namely, the computational substrate created within the computer that connects to it. So the BIV is *not* envatted, if by 'envatted' one means 'disconnected from the world'. It is also *false* that it is disembodied, if by 'embodied' one means having a neuro-informationally connected peripheral system, neuromuscular joints and muscles—all these exist in the computer, but, of course, they are not materially the same as our peripheral nerves and muscles.

However, here is a novel point to be made, which reflects the concern with the PNS that I have been pushing so far. Even

Chalmers's idea of the BIV as a metaphysical hypothesis neglects the PNS, which in the BIV case would be some cables (metal wires, optic fiber, laser, magnetic fields, or what not, *but still cables* in a generic sense), connecting the BIV to its computer. The full truth about the BIV lies with the cables. For the sake of simplicity and familiarity with the subject, let us consider the case of pain as experienced by the BIV. We want to simulate with the help of the computer that the BIV has a body, and because of peripheral nerve damage, say, of A-delta fibers, she is hypersensitive to touch on the skin of the feet. Now, in order to exactly match what is going on in such a nervous system, and so to *stimulate the brain in the right way*, the computer will have to *simulate* (i.e., to represent to itself), a mechanism like the Gate Control Theory posits to be present in the *substantia gelatinosa* of the spinal cord. No stimulation without representation!

The computer will have such a gate control mechanism, which will mediate efferent and afferent impulses of the BIV, and therefore simulate the whole pattern of neural firings that is normally taking place in subjects with peripheral neuropathy. So the computer will contain the relevant structures and function of the spinal cord, as well as the relevant PNS structures, except these will be implemented by artificial circuits, not neurons. It will have to do all this; otherwise it can't stimulate in the required orderly way the experience of pain. But wait a minute! This means that the computer doesn't merely *simulate* nervous structure in order to stimulate the BIV, but rather materially realizes, implements, or emulates it. It creates whatever is needed for the pain process to actually take place, and it is part of this process. More importantly, 'the cables', the PNS, is the BIV's PNS, but also the computer's PNS. The only difference is that during the process of pain experiencing, the BIV's afferent impulses are the computer's efferent impulses, and the BIV's efferent impulses are the computer's afferent ones. This is so since the pain process involves sensorimotor control mechanisms (i.e., sensation and action are indissolubly connected). This means that each has the other both as its body and as its brain; the computer is body for BIV, and the BIV is body for the computer. Each of them is both body and brain at the same time.

For instance, suppose that the computer has to 'touch the skin on the foot of the PNS-damaged BIV'. For instance, the larger perceptual context is that the BIV has the illusion of lying on a bed, and his cat, wanting to play, pokes his foot. In order to simulate this, the computer must really implement *some kind of skin* with peripheral

nerve terminations that are sensitive to touch, it must have *a kind of* damaged state of *some kind of* A-delta fibers, it must have *some kind of* C fibers, it must have a way to make the impulse travel at a slower pace (because of the nerve damage), and it must have *some kind of* a gate control mechanism in *some kind of* spinal cord, so that the BIV's cortex can receive the right nervous signals, and release its output via the action module to the motor nerves, via the cables that connect back to the computer, which motor signals the computer will now interpret as input, and the process can start from the beginning according to how the motor response is supposed to affect the relation between the computer-skin and the initial stimulus. My point is really simple: the BIV has so far been presented in the literature (for all I know) as a passive receiver of stimulation from the computer, when, in fact, they should be thought of as *interacting*. The motor signals of the BIV will require from the computer to act as well, so as to arrange and structure its stimulations to match the new situation created by the BIV's action. This is the phenomenon of *reafference*, whereby sensory signals result from the subjects own movements. For instance, suppose the BIV is lying in bed on his back and turns on his right side. This motor action will have to have the effect that a different part of the BIV's skin will be exposed to the pressure of the bed and to the wrinkles of the bed sheet. The computer will receive the motor signal from the BIV as an afferent signal, that is, ultimately as a sensory signal, based on which it will be able to create the right sensory signal for the BIV (viz., new tactile sensations, on a different region of the skin, etc.) and send this newly created sensory signal to the BIV. Sending this newly created signal is a motor action from the point of view of the BIV.[2]

So what are we to make of this elucidation? Is a BIV duplicate of yours phenomenally conscious? Is a BIV conscious? Does it have intentionality? Is it possible to create a BIV? All these questions have been answered positively or negatively, according as whether the philosophers in question are internalists about mental content and/or phenomenology or externalists. Some philosophers connected to the EE movement are more cautions and claim that

[2]An anonymous referee expressed disagreement with my idea that in fact the computer will have to simulate experiences in order to stimulate the BIV, because no one in the literature on the BIV thinks of the computer as simulating but merely as stimulating. However, that is precisely my point, namely, that this necessity to ultimately simulate in order to stimulate, and thus to re-create the whole PNS in the computer in some form or other, has been overlooked in the literature.

whether experiences can be simulated in a BIV is an open and empirical question. But if my reasoning above is correct, we can actually rule out that the BIV—supposed to have consciousness, perceptual and sensory states—is possible. The BIV and its connected computer are like two mirrors facing each other; there is no genuine information in the compound system. The electric nerve impulses are embedded in electric nerve impulses, which in turn are embedded in electric impulses again, not in anything like a world, or reality. This is true even if we accept Chalmers's point that the BIV is a metaphysical scenario. To see this, remember the discussion of the GCT approach to pain mechanisms. Melzack and Wall pointed out that *there is* a conceptual connection between a nerve fiber type being nociceptive and its being a pain-specialized fiber; that's how these fibers actually got their identity conditions postulated by neuroscientists. Yet, in our thought experiment with the BIV pain there is no stimulus that we can properly consider as *noxious* (or thermal, or tactile, or whatever). For that you need a world, in the sense of an *extra-neural reality*. The computer does not bring about such a reality; computer and BIV are caught in an infinite recursive reflection of meaningless electric signals.

II. Functionalist Troubles?

Functionalism based on the idea of understanding mental states as causal roles or causal role fillers is widely and implicitly assumed in the philosophy of mind. The quotations I have provided so far all indicate the assumption that experience is something caused by stimulation and causing motor response. One could find hundreds of other quotes that indicate this assumption. The problem is not with causation or causal role as such, but with the specific place mental states are posited as occupying in a causal chain. What causes trouble for functionalism is that mental states are taken as brain states occupying a role between sensory stimulation and motor response. Ordinarily, philosophers understand functionalism as stating that mental states are states caused by stimuli and, together with other mental states, cause behavior. What is not made explicit, but is implicitly present in how philosophers understand functionalism, is the assumption or prejudice against the PNS being part of mental states; the assumption is that mental states are states of the brain caused by external stimuli *and* excitations in the afferent PNS and causing excitations in efferent PNS and behavior.

If instead we adopt a different picture, one according to which the PNS is no less part of mental states than the CNS, some troubles of functionalism can easily be avoided. According to the new picture, it is still true that mental states are states caused by stimuli and causing behavior, but 'stimulus' and 'behavior' are understood differently. Stimulus is understood as an event external to the nervous system (e.g., a burn of the skin, a gallbladder stone stuck in the mucosal fold, light hitting the surface of the eye, etc.). Behavior is understood as a bodily event occurring posterior to the neuromuscular joint, hence outside the nervous system (e.g., a contraction of the biceps muscle, a contraction of the gallbladder's *muscularis* layer, a motion of the eyeball). Mental states are then states occupying the causal role between such kinds of events. I will briefly present the solutions to three problems with old functionalism, based on this new understanding of causal role.

replace with "junction"

replace with "four"

A. The Mad Pain Problem

David Lewis combines an analytic version (or commonsense) causal role functionalism with the identity thesis, according to which to be in pain means to be in a state with the pain-role, which by actual empirical identification will entail that to be in pain is to be in a certain brain state. Lewis (1980) recognizes that what he calls 'mad pain' is a problem for the functionalist part of the theory, assuming the pain-brain state identification is correct:

> There might be a strange man who sometimes feels pain, just as we do, but whose pain differs greatly from ours in its causes and effects. Our pain is typically caused by cuts, burns, pressure, and the like; his is caused by moderate exercise on an empty stomach. Our pain is generally distracting; his turns his mind to mathematics. . . . In short, he feels pain but his pain does not at all occupy the typical causal role of pain. . . . [M]y opinion that this is a possible case seems pretty firm. If I want a credible theory of mind, I need a theory that does not deny the possibility of mad pain. ([1980] 2000: 110)

Mad pain is indeed conceivable if, as Lewis presupposes, pain is a brain state merely causally connected to the PNS excitation, rather than being partly constituted by these. Lewis prefers to keep the idea of pain as a narrowly understood neural state, to mean 'brain state', and turns the initial functionalist definition of pain into a population-relativized version, according to which all we can properly define is 'pain-relative-to-a-population'. So the mad pain is just normal pain relative to the mad population, because relative to that

population the mad causal role is the normal one. Lewis move is ad-hoc and the resulting theory very inelegant; no wonder it has never become popular among functionalists.

The solution to the mad pain problem is simple. Mad pain is a logical impossibility. Pain is not a brain state or even a CNS state, but a global state of the entire nervous system. The PNS activities are partly constitutive of the state of pain, hence the strange man Lewis describes is not in pain at all, because moderate exercise on an empty stomach does not induce excitation in nociceptive peripheral nerves (whatever nerves are acted on by moderate exercise, they don't deserve the name "nociceptive fibers"), and because the motor response of, say, turning one's body, gaze, and so on toward some mathematical formula written on a paper is not a pain-specific motor response (whatever motor nerves are activated, they don't deserve the name "pain-specific motor peripheral nerves").

B. The Problem of Pseudo-Normal Vision

A classic argument against functionalism is the conceivability of color spectrum inverted pairs of people, that is people who are functionally identical, but have their phenomenal color experiences spectrum inverted; for instance, they both respond to the same color stimulus, say, red, in the same way, but one experiences red phenomenally, the other experiences green. Some functionalists might try to show that there is some deep logical incoherence at play. However, Martine Nida-Rümelin ([1996]) argued that spectrum inversion is to be taken seriously even empirically, as there are cases of inherited vision defects that seem to point to the *actuality* of inverted people.

Nida-Rümelin's case is that of pseudo-normal vision. There are three types of photoreceptors, called 'cone cells', on the retina that play a role in human color vision (in bright light conditions): R-, G-, and B-cones (from red, green, and blue).[3] They are morphologically distinguishable and normally each of them contains different photopigments, which absorb certain wavelengths of the incoming light, so that after this filtering the output nerve signal that is transmitted to the optic nerve will normally be different for different perceived colors. What colors are perceived is determined by the interaction among these three cone types. Red-green color blind

[3] Hence, human color perception is trichromatic (i.e., based on three basic colors: red, green, and blue). However, there are several studies that indicate there might be tetrachromat humans as well.

subjects (i.e., those who can't distinguish red and green) have their R- and G- cones containing the same photopigments. However, there are two types of such partial color blindness: (A) when the green photopigment is contained in both the G and the R cones, and (B) when the red photopigment is contained in both the G and R cones. The genes responsible for case (A) and (B) can in rare cases be present in one and the same person, thus such persons, called 'pseudo-normal', although not visually defective in terms of normal color discriminations, have their photopigments swapped between the G and R cones. Hence, according to Nida-Rümelin, they are actual cases of spectrum inversion.

But why does Nida-Rümelin think that these people are spectrum-inverted rather than just normal subjects, but whose red and green experiences are realized by different nervous system structures, a difference based on the G-cones, rather than R-cones, being involved in red experiences, and R-cones, rather than G-cones, being involved in green experiences? If what I have been arguing so far is right, then the PNS parts involved in color vision—in our case the G- and R-cones and the photopigments they contain—play a *constitutive* role in color vision. If I am right, G-cones get their name 'G-cone' in virtue of containing the green-sensitive photopigments; *mutatis mutandis* for R-cones. The point is the same as the one I made in connection with nociceptors: they deserve the name to the extent that they respond to noxious stimuli. Nida-Rümelin's idea that pseudo-normal vision is spectrum-inversion is precisely based on assuming that the PNS components are merely *causing* red or green experiences, the experience is a point *terminus* in the CNS, 'unaware of' what is happening at the level of G-cones and R-cones:

> the proposal violates the widely accepted principle of supervenience for mental properties upon the relevant physiological properties. Since the neural hardware is not affected by exchanging photopigments, we must assume that the physiological state produced by a specific pattern of stimulation of concrete photoreceptors in a given person is the same regardless of whether the photopigments are reversed. . . . [It] entails the prediction that the *very same* physiological state will lead to a red-sensation in the one case and to a green-sensation in the other. Since the only difference between the two cases lies in the way the physiological state is *caused* (by different patterns of light stimuli) and since the brain does not have any access to this information, this would seem rather mysterious. (emphasis in original)

Of course, Nida-Rümelin is right against the functionalist to the extent that both of them assume that the experience is supposed to

be brain-bound, hence supervene on the brain states—that's what she means in the above quote by 'relevant physiological properties' and 'neural hardware'. *If the old functionalist agrees that when I experience green, the pseudo-normal person experiences red*, then the old functionalist is committed to the lack of supervenience of experience on what the functionalist takes as the relevant physiological state, namely, a brain state. The above 'if', however, becomes a big 'if', in the context of our approach.

My proposal is that 'relevant physiological properties' and 'neural hardware' include the PNS. To consider an example, I and a pseudo-normal person are presented with a green object. What happens is that the G-cones on my retina are activated *in virtue of containing G-photopigment*. The R-cones in the pseudonormal person are activated *in virtue of containing G-photopigment*. Both of us see the object as green. Hence, there are no pseudo-normal people: both normal and 'pseudo-normal' subjects are functionally and experientially identical.

Nida-Rümelin is actually aware of such a response, when she says that a functionalist might respond that in a normal person who *becomes* pseudo-normal

> [T]hose individual receptors that were R-cones before the inversion of photopigment distribution in the retina of the person at issue, *turned into* G-cones. (1996: 104) (emphasis in original)

She finds this 'unacceptable' (p. 103), because "color vision science predicts such a person will experience and report a radical change in his color perception" (p. 104). Now, the truth is that vision science would in such a case predict verbal reports of radical changes in color perception only if the morphological differences between R-cones and G-cones make a difference to whether they are really G-cones and R-cones, *as far as their contribution to experience is concerned*. If being a G-cone as far as experiences are concerned is essentially being a G-cone as far as morphology is concerned, then a G-cone turning into an R-cone (to the extent that it can turn into one) will result in radical differences in experience. But the truth is that morphology does not play an essential role in the nature of the G-cones or R-cones as far as their role in experiences is concerned. What *does* play the only essential role is the photopigment; so a photoreceptive cell deserves the name 'G-cone' in virtue of containing the G pigment. By far the best proofs for this claim are two facts about research in morphological differentiation of photoreceptor cones:

- The relevant types of cones had been postulated and named about 100 years before any method to morphologically differentiate them became available.[4]
- Even to this day, there are no reliable methods to morphologically distinguish the cone types without first distinguishing them in terms of the light-pigment interactions.[5]

C. The China-Brain Problem

Ned Block's famous China-brain thought experiment ([1978] 2002) is supposed to boost the intuition that functional duplication of a phenomenally conscious system does not necessarily amount to phenomenal duplication of that system. Let me first quote a passage from Block:

> Imagine a body externally like a human body, say yours, but internally quite different. . . . Suppose we convert the government of China to functionalism, and we convince its officials that it would enormously enhance their international prestige to realize a human mind for an hour. We provide each of the billion people in China (I chose China because it has a billion inhabitants) with a specially designed two-way radio that connects them in the appropriate way to other persons and to the artificial body mentioned in the previous example. . . . Surely such a system is not physically impossible. It could be functionally equivalent to you for a short time, say an hour. ([1978] 2002: 96)

Block then claims that although the China-brain system is functionally equivalent to you, one might still coherently doubt that the system is conscious in the phenomenal sense. So, phenomenal consciousness is not necessitated by functional facts.

I think that why most people got moved by this thought experiment is because they focused their attention on the wrong side of the China system, namely on the brain. Block says that the govern-

[4] In 1802 Thomas Young postulated the existence of three types of photoreceptors in the eye, each responsible for detecting different ranges of wavelengths in the visible spectrum. The theory of trichromatic vision was developed in 1860 by Hermann von Helmholtz, in his *Handbuch der physiologischen Optik*, and this is the time when the R-, G-, and B- cones get their name. See Cahan 1993, part I. The existence of the cones was shown later, in the 1950s.

[5] The composition of cells is frequently obtained by light-dependent histochemical staining of the optically intact, or freshly excised eye, or of isolated retina *in vitro*. Alternatively, laser interferometry on the intact eye is performed. So the availability of morphological data that differentiate the cone types depends on interactions at the level of their pigment in the first place. See, for instance, Dacey and Lee (1994) for in vitro histochemical staining, and Roorda and Williams (1999) for laser interferometry.

ment of China can realize a human mind for an hour; I agree, but they can realize the human mind only because the CNS system they create is still connected to my body, to my PNS, that is.

I propose a thought experiment within this thought experiment. Suppose the government of China is a bit less ambitious and is content with only realizing a brief pain sensation via a system of people and radio transmitters. What happens is that I offer my nociceptive nerves and the corresponding motor nerves to be used by China for the experiment. They will put me to sleep, disconnect my pain-related PNS components from my brain and connect them to the China-pain CNS system. I will wake up, suppose, with all the other cognitive components intact, except for those involved in the sensation of pain. Intuitively, I will be zombified, pain-wise: I won't feel pain; yet, when my skin is hurt, my relevant muscles will contract and I will avoid the stimulus, I will scream, and so on.

What about the CNS states in the China-pain system? If in Block's experiment you intuited that the China-brain system is not conscious, you will intuit here that the China-pain system is not in pain.

I also intuit that in this case I am not in pain and the China-pain system is not in pain either. Yet, I do not intuit that there is no pain at all instantiated by the composite system of my body (PNS) + China-pain (CNS). To boost this intuition, consider a little change to the story. My PNS components are disconnected from my brain, just as before. But they are now connected to another person's pain-related CNS components; call that person 'John'. John's pain-related PNS components are disconnected from his brain, and left unconnected to anything. Suppose there is an intense mechanical action of a nasty stone in my gallbladder. As a result, my body bends, and I scream. But, as I said earlier, intuitively I'm not in pain, since I lack the CNS component of pain. The CNS component of this process is to be found in John's brain. Now, if we reconsider Melzack and Wall's assertion to the effect that 'the thalamus, the limbic system, the hypothalamus, the brain-stem reticular formation, the parietal cortex, and the frontal cortex are all implicated in pain perception', then all these components of the pain process are now in John's brain. These structures are responsible for awareness of one's pain.

So is John feeling pain? I think the only puzzle here is whether what John is aware of as pain is *his* pain, or *my* pain, or *no one's* pain. In other words, the only puzzle that arises at this point is related to the question of *who* is in pain, not *whether* something is

in pain, or whether there is pain somewhere in the global system <JOHN + ME>. *Who*-puzzles are puzzles to the extent that we take persons or mental entities seriously. But if we go back to our folk neuroscience and focus on nervous systems as subjects of pain, then my puzzle with John is less of a puzzle. We could call my nervous system S_i and John's S_j. We *do* intuit that neither S_i, nor S_j is unquestionably in pain, but only because we are not sure what to consider as part of S_i and S_j, respectively. Mereologically, the CNS component of the pain is in S_j, but in terms of neural networks it is both in S_i and in S_j. But all this indecision on how to demarcate S_i from S_j when it comes to the pain sensation can be resolved if we define a third system, S_k, as the neural network containing my pain-related PNS components, PNS_i, and John's connected CNS component, CNS_j. My intuition is that just because this system transcends the normal boundaries of nervous systems in living organisms, it doesn't mean that it does not instantiate, or *might* fail instantiate pain. It is active in the right way, so it is in pain. To say otherwise is to say that a colorful image contained in a JPG file on my hard disk becomes black-and-white just because it is transferred to an external hard disk. And to insist that, still, even if the system is in pain, it is not, or *might* not be in *phenomenal* pain is, again, to commit oneself to a funny notion of phenomenal pain as having nothing to do with ordinary pain. More importantly, to appeal to primitive phenomenality intuitions at this point, when the very question of phenomenality depends on what to say about the China-pain system in light of what we say about the John-me system, is tantamount to begging the question.

Now, the case with the China-pain system is no different from the case with John's CNS pain system. So the only puzzle about the original China-brain system is a puzzle about *who* is conscious, not about *whether* there is any consciousness somewhere in the global <MY BODY + CHINA-BRAIN> system.

D. The Triviality Problem

One of the arguments against functionalism about mental states is that it is a trivial claim, so that functional organization does not distinguish the mind from intuitively nonmental entities. The thesis is sometimes put as follows: even a bucket of water sitting in the sun is causally complex enough for there to be an interpretation of its states that would correspond to a realization of a human mind (Godfrey-Smith 2009: 273–74). To put it differently, since according to functionalism mental states are to be individuated by the their

place in an abstract causal structure including inputs, outputs, and internal states, there is no reason to think that only brains could serve this purpose, but any physical system that is complex enough to be interpreted as having all the relevant possible functional states and the rules that govern their occurrence, as specified by a state transition matrix. If the world contains many such physical systems that are not intuitively minds, then functionalism does not say anything interesting about the mind, as it cannot distinguish it from any complex enough systems. Several authors have offered various such triviality arguments against functionalism. I will focus here on the most recent such argument, by Godfrey-Smith, which is supposed to improve upon earlier versions. He offers a new triviality argument based precisely on the idea that what he calls "transducer layer," that is, the far side fringes of the PNS, and actually the whole PNS as such, was not taken into consideration by earlier arguments. The argument is that any structurally and causally complex enough brainlike physical system (what Godfrey-Smith calls "control system," and what we more naturally call a CNS) could be a mind as far as functionalism is concerned, because adding a relevant PNS to it does not change some of its mental properties, given that it is plausible to think that some such mental properties are indeed PNS-independent. The crucial premise here is, of course, that there is mentality without or independent of the PNS, and denying and considering this premise a prejudice is precisely what this book is about. Here is a quote from Godfrey-Smith that exhibits this prejudice:

> A bucket of water cannot possibly have the same functional profile as a human agent, as it does not have the right input–output properties. But we now look at the possibility of taking a functionally characterized system and *changing* its transducer layer, while keeping the control system intact. This is done by changing the physical devices that interface with external objects. We might alter the hair cells in the ear so they are not moved by vibrations, but by magnetic fields. We might have muscle fibers moving a mouse on a computer screen. Altering transducer layers has important therapeutic possibilities for people with sensory and motor disabilities.
>
> When the transducer layer of an intelligent system is altered, what are the consequences for its psychological properties? . . . There may be many psychological changes implied, but it is natural to think there are *some* mental features of an agent that depend only on the properties of the control system, and are unaffected by the properties of the transducer layer. . . .

> But if a system has non-marginal mental properties, a mere change to its transducer layer should not alter this fact. Two functionally similar systems that differ only in physical make-up and transducer layer must either both have, or both not have, non-marginal mental properties. So if the bucket of water lacks only the right transducer layer to be a functional duplicate of A [N.B. *a human agent*], then it must already have some non-marginal mental properties. (2009: 285–87)

Godfrey-Smith's talk about "non-marginal mental properties"—by which he means mental properties that would not be had by, say, a worm, because an alteration to its PNS would essentially alter its mind—is the perfect symptom of CNS-fixation that is so widespread in contemporary philosophy of mind. Why assume that the mind does not essentially and constitutively include the PNS, the transducer layer, to use Godfrey-Smith's terminology? Why assume that there is a difference in kind between our minds and worms'?[6] Be that as it may, once we take the PNS seriously, the solution to the triviality problem is quite simple and elegant. The "margins of the mind" (i.e., the PNS) are as much part of the mind as the so-called "central mind" (i.e., the CNS or the brain).

The reason the bucket of water is not conscious is precisely because, however complex it is internally, it lacks a PNS that would lawfully respond to stimuli, connect to the water's internal states, and be caused by the latter to set an effector in motion, say, a musclelike tissue. If we added such a PNS to the bucket, it *would* indeed be conscious. Having the right PNS-like structure is what makes a relevant complex system count as a *nervous system*. Suppose we added to the micro-structurally sufficiently complex quantity of water in the bucket a PNS-like sensory-motor anatomical structure. For that matter, let's connect two anatomically complete human arms, with intact nociceptive nerve fibers and intact motor fibers to

[6] The brain-fixation prejudice in the philosophy of mind is also, in my opinion, the ground for another widely held prejudice; the view that phenomenal states are only instantiated by higher animals. Virtually all contemporary philosophers claim, without any argument, that dogs can certainly feel pain, but "simple" organisms like worms or slugs most probably do not have phenomenal states whatsoever. Now, to think that worms and slugs are neurologically simple is another blunder of contemporary, scientifically uninformed philosophy. To take as an example the current "superstar" nematode worm—superstar, because it was the first multicellular organism to have its genome completely sequenced, by 1998, and is widely used as a model organism—the 1 mm long *Caenorhabditis elegans* exhibits a nervous system of 302 neurons and a sensorimotor system with very complex connectivity patterns. For a dynamic interactive online visualization of these connections within the worm's neural network, see http://wormweb.org/.

the bucket of water, in such a way that whenever we effect some damage to the skin of the hands, the hands will move away, in specific ways, from the source of the noxious stimulus. Does the hand feel pain? Well, it responds to painful stimuli; it has the specific specialized sensors. So, of course, it does. Does it mean that the bucket feels pain? Not exactly, but close: the <bucket + PNS> nervous system *is indeed in pain*. The whole intuition that buckets of water sitting in the sun are not minds is present only to the extent that a properly connected PNS is lacking, which makes the bucket itself not count as a nervous system. But the same is true for the brain: were it not properly connected to a PNS, it would not count as a mind. The triviality arguments work against the old functionalism, which is based on the brain-fixation prejudice, but, as I have argued, the new CNS + PNS involving functionalism is a very plausible theory and it is immune to the triviality objection.