

## **Can neural networks help us explain the phenomena of consciousness?**

Domenico Parisi

Institute of Psychology  
National Research Council, Rome  
parisi@ip.rm.cnr.it

### **1. Farewell to consciousness**

If we want to understand what is consciousness the first thing we should do is avoid using the word. The word “consciousness” is full of ambiguities, vagueness, pseudo-problems that can only be eliminated if we simply avoid using it. Another problem with the word “consciousness” is that this word suggests that there is a single, unified entity called “consciousness”, whereas the expression is used to refer to a variety of different phenomena that should be examined and understood separately. Consciousness is like the sun which if looked at directly dazzles us and prevents us from seeing anything. Consciousness should be studied by examining separately and in detail the different empirical phenomena we refer to by using the word “consciousness”. In the present work we will not use the word “consciousness” (except in the title and in the present section) although we will be concerned with phenomena that are frequently referred to by using the word.

The second thing that science should do if it wants to understand the phenomena of consciousness is adopt neural networks as theoretical models to interpret and explain these phenomena. Neural models are simulation models of behavior and mental life that are directly inspired by the physical structure and way of functioning of the nervous system. Simulation models are theoretical models which are expressed as computer programs. Theoretical models expressed as computer programs, i.e., simulations, are necessarily explicit, internally complete, and detailed because, otherwise, the program would not “run” in the computer. Furthermore, simulations necessarily make a large number of detailed empirical predictions since the results of a simulation are the empirical predictions derived from the theory incorporated in the simulation. The phenomena of consciousness are very complicated and elusive phenomena and therefore it should be advantageous to interpret them using theoretical models that have these properties instead of using purely verbal models, as is mostly done. Another advantage of using neural networks is that neural networks link up the study of consciousness with the study of the nervous system, a connection which is particularly useful given the rapid and constant progresses which are made today in the study of the nervous system, including those aspects of the nervous system that underly the phenomena of consciousness (Edelman and Tononi, 2000; Metzinger, 2000; Dehaene, in press). Neural networks use the same conceptual vocabulary of the natural sciences as they only speak of processes in which physical causes produce physical effects and everything has an intrinsically quantitative nature. Therefore, they make it possible to integrate the study of consciousness phenomena in the theoretical framework of the natural sciences.

In the present work we interpret various phenomena and aspects of consciousness using neural network models. However, the type of neural networks that we will use are not “classical” neural networks. The neural networks that we think are more appropriate to interpret consciousness phenomena (and, we believe, more appropriate in general) are neural networks viewed in an

Artificial Life perspective. In Section 2 we briefly discuss the distinction between classical neural networks and neural networks viewed in an Artificial Life perspective and in the sections that follow we examine various consciousness-related phenomena using the latter type of neural network models. As we have said, we will not use the term consciousness but we will speak of the distinction between public and private reality (Section 3), mental life as distinct from behavior (Section 4), the distinction between oneself and the rest of reality (Section 5), feeling and the two different types of action that from outside can influence the functioning of a neural network (Section 6), and subjective experience (Section 7). In the final Section 8 we will try to answer the question: What is the difference between a person who sees that the room is lighted or dark and a photoelectric cell that makes the same type of distinction? To answer that the difference is that the person has consciousness and the photoelectric cell does not, is not very useful. Instead, the models of the various phenomena we deal with in Sections 3-7 may help us to give more useful answers and remain within the domain of the natural sciences.

## **2. Neural networks in an Artificial Life perspective**

Classical neural networks are the neural networks introduced in Hopfield (1982) and Rumelhart and McClelland (1986) and used most frequently in the neural network literature. All neural networks are sets of neuron-like units connected by synapse-like connections. Units have activation levels which, for input units, depend on events outside the network and, for internal and output units, depend on excitations and inhibitions arriving to each unit from connected units. From the point of view of studying consciousness phenomena the two crucial differences between classical neural networks and neural networks in an Artificial Life perspective are that, unlike classical neural networks, neural networks in an Artificial Life perspective have a physical body and live in a physical environment. Classical networks tend to be viewed as abstract “information processing devices”. The researcher decides which input arrives to the network at any given time and the network is trained to respond to the input with the appropriate output. In contrast, neural networks in an Artificial Life perspective are models of a physical object, the nervous system, which is physically contained in a larger physical object, the organism’s body, and the organism’s body is contained in a still larger object, the physical environment in which the organism lives and with which it interacts. What is of interest and what is simulated in Artificial Life simulations is not (only) what takes place inside the neural network but the interactions between the neural network and the rest of the organism’s body, on one side, and the external environment, on the other (Parisi, Cecconi, and Nolfi, 1990; Nolfi and Floreano, 2000; Parisi, in press).

Two consequences of this view of neural networks are particularly important from the point of view of consciousness phenomena. First, the inputs that arrive to a neural network have different properties depending on whether they arrive from the external environment, from inside the organism’s own body, or from inside the neural network itself. Second, unlike classical networks, neural networks in an Artificial Life perspective have some control on their own input. As we have already noted, in classical neural networks it is the researcher who decides which input is presented to the network at any given time (i.e., in any input-output cycle) and it is the researcher who reacts to the output with which the network responds to the input (for example, the researcher evaluates the output and provides a teaching input to the network). On the contrary, neural networks in an Artificial Life perspective can change with their output the external environment or they can change the physical relation of their body or body parts to the external environment, thereby influencing the future inputs they will receive from the external environment. Or the output of a neural network can have as its consequence a change in the state of the rest of the organism’s body, and therefore the network’s output will influence the future inputs that will arrive to the network from inside the organism’s body. An important consequence of this control that neural networks can have on their own inputs, because they change with their output the state of the sources of their inputs, is that

neural networks can (learn to) predict the consequences of their actions. The network's ability to predict the consequences of its own actions, that is, to predict the next input given both the current input and a planned action, varies with the type the input, i.e., with whether the input comes from the external environment, from inside the network's own body, or from the inside the network itself, and, in the case of input arriving from the external environment, with whether the input results from the output (behavior) of another network (a conspecific) or from other objects and features of the external environment.

### **3. Public reality and private reality**

A human being lives in two different realities, a private reality and a public reality. An individual's private reality is what the individual feels, his/her mental images, memories (rememberings), thoughts, predictions, plans, desires and fears. This reality is private in that it is only accessible to the individual and to nobody else. The individual can communicate the content of his/her private reality to other individuals by using the words of language or other communicative signals such as his/her facial expressions but in this case the manner in which the other individual has experience of the first individual's private reality is very different from the first individual's own experience. In contrast, public reality is the world of things and events to which everyone has in principle the same type of access and which therefore is shared by each individual with all the other individuals. There is only one public reality whereas there are as many private realities as there are individuals.

The private reality of persons constitutes a problem for science since science is of public reality, of what can be observed by many individuals and can be measured by comparing it with publicly available measuring instruments, and science finds it difficult to deal with something which is only accessible to the single individual. However, the task of science is to study all of reality and to connect all aspects of reality together, and therefore science cannot ignore the private reality of persons. Of course, science cannot but study private reality in its own characteristic way, that is, publicly and by connecting private reality to public reality. Therefore, the first problem for science when it confronts private reality is to explain how it is possible that human beings live in two different realities, a private and a public reality. This is the problem we tackle in this section by using neural network models.

As we have said, neural networks are theoretical models that interpret the behavior of organisms, and their mental life (when they have one), in terms that are directly inspired by the physical structure and way of functioning of the nervous system. A neural network is formed by a certain number of units (neurons) each of which has, at any given time, a quantitative level of activation (firing rate of neurons). The activation level of the network's input units is determined by physical or chemical events that take place just outside the network. The input units are connected by unidirectional connections (synapses between neurons) with the internal units and the internal units are connected with the output units, so that activation spreads through the network of connections. Behavior is the way in which the network responds (pattern of activation of the network's output units) to the input given the weight (number of synaptic sites between neurons) and "sign" (excitatory and inhibitory) of each different connection.

A neural network is a simplified model of the nervous system of an organism, that is, a model of that particular bodily organ or system which is specialized for the control of the organism's behavior. But the organism's body does not contain only the nervous system. It contains many other organs and systems that cannot be ignored if we want to explain the organism's behavior.

As we have said, the pattern of activation of a neural network's input units is caused by particular physical and chemical events taking place outside the network. To explain how it is possible that a

human being lives in two different realities, one public and the other private, we have to consider that the physical-chemical events that results in the patterns of activation of the network's input units can take place either in the environment outside the organism's body or inside the organism's own body. In the first case these causes are light or sound waves, mechanical pressures, or chemical molecules released by substances smelled or tasted. For purely physical reasons these causes can produce more or less similar patterns of activation in the input of units of the neural network (nervous system) of many different individuals, provided the individuals are located sufficiently close, and therefore they can influence in at least initially similar ways what takes place inside the various individuals' neural networks. The reality revealed by these activation patterns will therefore be a public reality, a reality accessible to many different individuals. The light that falls on an object located in the external environment and is reflected by the object will arrive to the input units (visual receptors) of all the individuals that are sufficiently close to the object and will cause approximately the same visual experience in all these individuals. The pressure waves caused by the vibration of an elastic object will arrive to the input units (acoustic receptors) of many different individuals and will cause the same auditory experience in all of them. The molecules coming out of an open perfume bottle and diffusing in the air can reach the input units (olfactory receptors) of many different individuals causing the same olfactory experience in all of them.

The public reality revealed by the activation patterns that are the effects of this type of causes is not a reality of which the individuals can only have a passive experience. The various individuals can act on the objects out there, by moving seen objects, causing the vibration of objects in order to produce sounds, opening and closing perfume bottles, and as a result of these actions all of them will experience similar or systematically related consequences, similar changes in the activation patterns in their network's input units, and therefore in their experiences.

However, the activation patterns in the input units of an individual's neural network can also be the effects of events taking place not in the external environment, outside the organism's body, but inside an individual's own body. These events will cause activation patterns in the input units of the individual's neural network but, for purely physical reasons, they will not be able to produce similar activation patterns in the input units of the neural network of other individuals. When an individual moves his/her arm or hand, the input units located in the muscles of his/her arm or hand (kinesthetic receptors) will have specific activation patterns that are the effects of these movements but these effects won't be recorded in the kinesthetic receptors of any other individual. If the sugar level in the blood of an individual goes down below a certain threshold, specific molecules will bring a "hunger" message to the individual's neural network, but this will happen only for the particular individual, not for other individuals, even if they are close in space. The reality revealed by these activation patterns caused by events inside the individual's body will be a private reality, accessible only to the particular individual, and to nobody else.

Let us now turn our attention to the output of an individual's neural network. The activation patterns of the network's output units have physical or chemical effects outside the neural network itself. But these effects can take place both in the external environment or inside the organism's body. The output units' activation patterns can result in movements of the entire body of the organism or of particular body parts, in the production of sounds, in the opening of a perfume bottle, or they can result in some change of state inside the organism's body. These changes caused by the organism's neural network either in the external environment or inside its own body will tend to result in public inputs for the individual's neural network in the former case and in private inputs in the latter case.

We conclude that neural network models can help us explain how it is possible that individuals live both in a public reality shared with other individuals and in a private reality which is only accessible

to each particular individual. What is important in this account is that it is an account which invokes only physico-chemical causes producing physico-chemical effects and therefore remains entirely within the conceptual framework of the natural sciences.

Notice that this account does not imply that all organisms are able to distinguish between public and private reality. It does imply that all organisms, even the simplest ones, receive inputs from both the external environment and from inside their own body, and that inputs from the external environment can be inputs for a plurality of individuals whereas inputs from an individual's own body can be inputs only for that individual. But many simple or nonsocial organisms may not be able to make the distinction between public and private reality. We will see later in this chapter what else is needed to be an organism which, like human beings, can tell the difference. But we will first discuss, in the next section, a neural network account of mental life.

#### **4. Mental life**

Both the inputs that are caused by events in the external environment and the inputs caused by events inside the organism's body are effects resulting from something which takes place outside the individual's neural network. Something happens outside the neural network, either in the external environment or in inside the organism's body, and this something influences the neural network by producing a particular pattern of activation in the network's input units. The same for the output. The activation patterns in the network's output units have effects that take place outside the network itself, both in the external environment (the organism's body is moved, a sound is produced, a perfume bottle is opened) and inside the organism's body but outside the neural network.

The neural network of simple organisms is limited to these interactions with what is outside the neural network. But this not so for the neural network of more complex organisms such as human beings. The neural network of a human being is characterized by the fact that its activity does not result only or necessarily in effects that take place outside the neural network but it can result in effects that take place inside the neural network itself. Similarly, the neural network of human beings is not only influenced or triggered to action by inputs coming from outside the network, either from the external environment or from inside the organisms' body, but also by inputs which are generated inside the neural network itself (self-generated inputs). A neural network of this type is capable of having mental life.

To simulate (or to begin to simulate) mental life it is necessary to go from neural networks that only have "forward" connections to neural networks that also have "backward" connections. Forward connections are connections that propagate the activation from the input units to the internal units and then from the internal units to the output units. Backward connection, instead, connect layers of units that are near the output units with layers of units that are near the input units. Therefore, backward connections propagate the activation in the opposite direction with respect to forward connections. Simple neural networks, that do not support a mental life but only a behavior, do not have backward connections but only forward connections. More complex neural networks, that support not only behavior but also mental life, have a rich structure of backward connections.

Let us imagine a neural network with three layers of internal units intermediate between the input layer and the output layer. Let us call the internal layer which is closer to the input units "post-sensory layer" and the internal layer which is closer to the output layer "pre-motor layer". We call the internal layer intermediate between the post-sensory and the pre-motor layer "central layer". (Any reference to specific portions of the actual nervous system is only approximate and in any case one should not identify, for example, what we have called the pre-motor layer with what is called

pre-motor cortex in the actual brain.) If the neural network does not support mental life all the network connections are forward connections and activation is propagated from the input units to the three successive layers of internal units and then to the output units. On the contrary, if the neural network is to be capable of mental life, it must contain backward connections, and more specifically backward connections leading from the pre-motor layer to the post-sensory layer.

Let us imagine that some activation pattern is activated from outside in the network's input units. Activation is propagated from the input (sensory) units to the post-sensory layer, from there to the central layer and then to the pre-motor layer, until it reaches the output units. However, given the backward connections the activation can also be propagated in another way. When the activation reaches the pre-motor layer it can proceed further forward to the output units or it can be retro-propagated to the post-sensory layer through the backward connections. The pattern of activation in the post-sensory layer which results from this retro-propagation of the activation constitutes a kind of self-generated input for the neural network. Inputs from outside are those that appear in the sensory input layer, which constitutes the interface between the neural network and the external world. However, our network can also self-generate its own inputs. Self-generated inputs are not observed in the sensory input layer but in the internal layer immediately downstream with respect to the sensory input layer, i.e., in what we have called the post-sensory layer. The network can respond both to external input in the sensory input layer and to self-generated input in the post-sensory internal layer. When it responds to the latter type of input, it is living a mental life.

A person is in front of the Coliseum in Rome and is looking at the ancient building in daylight. The light waves that reach the Coliseum from the sun and are reflected back to the person's visual receptors in his/her retina produce an activation pattern in the visual input units of the person's neural network. This activation pattern is transformed into another activation pattern in the post-sensory layer by the connections leading from the input layer to the post-sensory layer. Let us assume that it is this activation pattern in the neural network's post-sensory layer which allows the individual to "see" the Coliseum. Let us imagine now that in another occasion the person is many kilometers away from the Coliseum. Given the distance he/she cannot "see" the Coliseum but let us assume that he/she is imagining the Coliseum, is having a visual mental image of the Coliseum. There are no light waves that are reflected back from the monument and reach the individual's retina. The individual is not "seeing" the Coliseum. However, inside the individual's neural network activation is propagated backward from the pre-motor layer to the post-sensory layer and this activation creates an activation pattern in the post-sensory layer which is similar, even if not identical, to the activation pattern triggered from outside which we have assumed explains the individual's actually seeing the Coliseum. The individual is "imagining" the Coliseum.

The activation which produces in the post-sensory layer the visual mental image of the Coliseum can have its origin in the external environment, for example in a verbally expressed command from another person ("Imagine the Coliseum"), or it can be the result of an endogenous activity inside the neural network. In both cases the mental image functions as a (self-generated) input because it can be the beginning of further activity inside the network. Mental images are a typical example of mental life. Other examples are rememberings, thoughts, predictions of future inputs, planned but nonexecuted actions. In all these cases an internal activity inside the neural network goes through the backward connections and produces an activation pattern in the post-sensory layer which functions as a self-generated input for the further activity of the neural network. This is the basic mechanism. Additional conditions distinguish among various types of mental life. Rememberings or memories are distinct from mental images because they tend to reproduce specific episodes and events experienced by the individual in the past. Thoughts generally involve language even if language has a significant role in all mental life in that quite often self-generated inputs are imagined linguistic inputs, that is, activation patterns in the post-sensory layer that are similar to the

activation patterns produced by linguistic sounds arriving from outside, from other people or from the individual's own phono-articulatory movements.

Other important manifestations of mental life are predictions of future inputs and planned actions. Predictions of future inputs are activation patterns in the post-sensory layer that the neural network learns to produce in such a way that these activation patterns match the activation patterns that will appear in the input layer at some later time. Given the input currently arriving to the network from outside, the network is able to self-generate an input in the post-sensory layer that corresponds to the next input that will arrive from outside. Predictions of future inputs can be of future inputs that will arrive to the network independently from how the network reacts to the current input (seeing a flash of lightning and predicting a thunder, predicting the next word in a sentence given the preceding words; cf. Elman, 1990) and of future inputs that depend on some action with which the network plans to react to the current input (planning to drop a glass and predicting that the glass will break). In the latter case the prediction is generated by the neural network not only in response to the current input but also in response to a currently planned but not yet physically executed action, where a planned action is a pattern of activation in the pre-motor layer which is inhibited from propagating to the motor output layer but is propagated to the post-sensory layer to serve as the basis for generating the prediction of what the external input will be when the action will be physically executed (Cecconi and Parisi, 1990; Nolfi, Elman, and Parisi, 1994).

Planned actions and predictions of their consequences can also be self-generated in order to self-generate, in turn, predictions of rewards or punishments that may result from these consequences in order to decide whether to actually execute the planned action or to refrain from doing so.

With the addition of backward connections many interesting things can happen inside a neural network that are not possible in neural networks with exclusively forward connections. External inputs can produce self-generated inputs with no further consequences, i.e., without any immediate external output: something we see or hear evokes in us an image, a thought, reminds us of something else, makes us generate a prediction or plan an action. Or a self-generated input can translate into some external output: a mental image, a thought, something we remember, leads us to engage in some action. Or everything can happen and remain inside the neural network: self-generated inputs produce other self-generated inputs in a continuing recirculation of the nervous activity. In human beings it seems as if mental life never stops. We rarely respond to external inputs with external outputs, and stop there. Everything we see or hear or smell or taste tends to have an accompaniment of mental images, rememberings, thoughts and, possibly, hypotheses of actions. Every action directed towards the external environment is first internally recycled to predict and evaluate its consequences.

Of course mental life is private like the reality revealed by the inputs that arrive to the neural network from inside the body. For the same obvious physical reasons, what causes self-generated inputs inside an individual's network can only cause those inputs in that individual's network and not in the neural network of any other individual. Mental life is part of the private reality of each person, not accessible, except through communication, to other individuals. But communicating a mental image, a thought, something remembered, is behavior, not mental life, and it is the behavior's effects which arrive to the network of other individuals, not the image itself, the thought, what was remembered.

Mental life comprises most of what is called "subjective experience", what is subjectively lived. But mental life is not the whole of experience because, as we define it here, mental life constitutes only the cognitive or intellectual aspect of experience. Experience includes other fundamental aspects which are not strictly cognitive or intellectual but are dynamical, motivational, emotional: what a

person feels, not what it imagines, remembers, or thinks. But these dynamical aspects of experience can only be captured by considering the interactions between the neural network and the rest of the organism's body. Mental life takes place inside an individual's neural network. The dynamical aspects of experience result from the interactions between the individual's neural network and the rest of the individual's body. We will come back to this in Section 6 below.

## 5. Oneself and the rest of reality

The neural network of all organisms, whatever their species, receives two types of inputs: inputs that can only be received by the single individual (inputs from inside the individual's body and, for organisms endowed with mental life, inputs self-generated inside the individual's neural network) and inputs that can be received by all individuals provided they are located sufficiently close to one another. However, as we have anticipated, this does not mean that all organisms live in two realities, a private reality made accessible by the first type of inputs and a public reality made accessible by the second type of inputs, in the sense that all organisms can distinguish between the two kinds of reality. In order to live in two worlds and to be able to distinguish between the two worlds one must be a cognitively and socially sophisticated organism. Simple organisms live in an undifferentiated reality, which is neither private nor public. Let us try to explain this.

The crucial ability which distinguishes cognitively and socially sophisticated organisms from simpler organisms is the ability to predict future inputs. As we have seen in the preceding section, this predictive ability requires a neural network with backward connections. Only organisms with a nervous system sufficiently complex to include a rich structure of backward connections allowing the organisms to self-generate predictions of future inputs can have that cognitive and social sophistication which allows them to divide the reality in which they live into two distinct realities: oneself and the rest of reality.

The ability to distinguish between oneself from the rest of reality emerges on the basis of at least three components that we will examine separately: the differentiation of one's body from the rest of physical reality, the emergence of a mental life as distinct from behavior, the differentiation of one's mind from the mind of other individuals.

### 5.1 One's body

A first aspect of the division of reality into oneself and the rest of reality consists in identifying a specific portion of physical reality as "one's body" as distinct from the rest of physical reality. This identification is made possible by certain differences that the organism notices when it makes predictions concerning its own body and the rest of physical reality. An example is the following. When the organism sees two physical objects that come into contact, two different things can happen. In some cases the organism can predict this contact just before it takes place (on the basis of the spatial trajectory of the two objects or of one of them), it can predict some noise that will result from the contact, it can predict that two objects will not penetrate into each other, and so on. However, in other circumstances the organism is able to make all these predictions but, furthermore, it can predict that from the contact a tactile or thermic sensation will also result. In the first case both objects belong to the rest of physical reality but in the second case one of the two objects (e.g., one's hand touching an object) or both objects (e.g. one's hand touching one's face) belong to one's body. In such a way the physical object which is one's body begins to be differentiated from the rest of physical reality and to assume a unique identity within physical reality.

Another difference concerning predictions that allows an organism to distinguish between one's body and the rest of physical reality is that in the case of predictions about the rest of physical reality (e.g., two objects coming into contact) the organism is able to make these predictions without taking its own planned actions into consideration (as self-generated inputs), whereas in the case of predictions that involve one's own body (e.g., one's hand coming into contact with an object) the organism notices that it is able to generate these predictions only by taking its own planned actions into consideration.

A further basis for distinguishing between one's body and the rest of physical reality is that the objects that make up the rest of physical reality can rotate in front of the organism or be rotated by the organism or the organism can move around them, in such a way that the organism can see all their sides, whereas this is not possible for the physical object which is one's body. As a consequence, the visual access to one's body is quite restricted (unless one uses mirrors). This is another way in which a portion of physical reality becomes "my body".

## **5.2 Mind differentiates from body**

The emergence of a "mind" (not necessarily "my mind") as distinct from physical reality (including both my body and the rest of physical reality) may result from processes and mechanisms such as the following. The organism discovers that some of its activities have consequences for the external environment in that the organism's neural network can predict these consequences. A physical action on the part of the organism has an influence on the inputs arriving to the organism from the external environment in that the organism's action modifies the physical relation of the organism's body or body parts to the external environment (this happens when the organism moves its eyes or its head) or modifies the external environment itself (this happens when the organism displaces or modifies an object). The organism can predict these changes in the inputs arriving from the external environment given a planned action and these predictions will be confirmed when the action will actually be executed. In contrast, an action which is planned but is not physically executed and therefore remains inside the neural network does not change the inputs from the external environment because the planned action does not modify the external environment or the organism's relation to the external environment. As a consequence, planned actions and their predicted effects begin to be part of a "mental reality" which is distinct from the "physical reality" to which physically executed actions and their consequences belong.

A "mind" can also emerge as the product of another mechanism. As we have said, physical actions have consequences on the external environment that the organism perceives as changes in the inputs arriving from the external environment. In contrast, as we have also said, planned but nonphysically executed actions do not have these consequences. However, the organism notices that all its mental activities, not only its planned actions but also its mental images, thoughts, and rememberings, while they do not have consequences with respect to the public inputs from the external environment, they do have consequences with respect to the private inputs arriving from inside its own body. This becomes a further criterion that the organism can use to isolate one portion of reality, or of its experience of reality, and make this portion of reality its "mind". The organism's mind influences directly the private inputs arriving to the organism's neural network from inside the body but the mind can influence the public inputs arriving from the external environment only if the mind becomes something different, i.e., physical actions executed by the body.

A further mechanism that creates a distinction between one's mind and one's body is the following. The mind has at the same time more and less control on the inputs arriving to the organism's neural network than the body has. On one side, mental images and thoughts have some resemblance to real objects and events perceived in the external environment but they are also somewhat different

because one is free to create and to manipulate mental images and thoughts in ways that are not possible with physical objects and state of affairs in the external environment. Hence, the organism seems to have more control on its own mental life than it has on the external environment. But, on the other side, the external environment is more stable and it can be more reliably controlled using physical actions such as approaching an object, moving around the object, permanently changing its properties, whereas the contents of mental life such as mental images, thoughts, rememberings, feelings, desires and fears, appear to be more elusive, volatile, and uncontrollable entities. (Art, that is, the creation of publicly available external artifacts that produce interesting feelings in the individuals who perceive them, can be interpreted as a way to make mental life less elusive and more controllable.)

### **5.3 One's own mind and the mind of others: intersubjectivity**

The mind that emerges from the processes and mechanisms we have described is not one's mind yet, that is, a mind which is distinct from the mind of other people. One's mind emerges as a result of a process of differentiation between one's mind and the mind of others. For this differentiation to take place, however, it is necessary that the organism be not only cognitively (i.e., endowed with sophisticated prediction abilities) but also socially sophisticated. What is a socially sophisticated organism?

A socially sophisticated organism is, first, an organism with a tendency to spend a large portion of its time physically close to other organisms of the same species and, second, an organism that interacts in sophisticated ways with these other organisms. The two things do not go necessarily together. There are animal species that are intrinsically nonsocial in that an individual spends most of its life in isolation from conspecifics (for example, most snakes) and there are animal species that, although they are social in the sense that many individuals live in close spatial association, have social interactions which are not necessarily very sophisticated (for example, many fish and insects). To be a social species it is not necessary to be a cognitively sophisticated organism, whereas only cognitively sophisticated organisms can have socially sophisticated interactions. Humans are socially sophisticated (have socially sophisticated interactions with conspecifics) because they are cognitively sophisticated.

What is a socially sophisticated interaction? In very general terms, organisms behave in certain ways because these ways of behaving cause certain effects. If their behaviors would not have those particular effects, they would not behave in that particular way. This applies to all organisms since the behaviors of all organisms are selected during evolution or during the lifetime of the individual because of their effects. What is specific of human beings are two factors that make human sociality very different from the sociality of other animals, including nonhuman primates. The first factor is the typically human capacity to (learn to) predict the consequences of their own actions. All social organisms, since they live in close contacts with conspecifics, with their behavior tend to produce effects in other individuals, and viceversa, the behavior of other individuals tend to have effects on them. This does imply that all social animals have the tendency/capacity to learn to predict these social effects. Human beings, with their nervous system which is specialized for learning to predict the consequences of their own actions, learn to predict the consequences that their behavior will have on other individuals and the effects that the behavior of other individuals will have for themselves. This is the first factor that is responsible for the fact that social interactions in humans are sophisticated.

The second factor that makes human social interactions more sophisticated and different from the social interactions of other animals is that humans, unlike other animals, tend to reach almost all their goals not individually but with the direct or indirect participation of other individuals. In

nonhuman animals an individual realizes almost all of its goals (which, incidentally, are not as numerous, complex, and differentiated as the goals of a typical human being even in simple cultures) with its own individual actions. (The only important exception to this rule is, in some animal species, caring for the newborn.) On the contrary, a human individual tends to need the help or the participation of other individuals in order to reach its goals. This creates a very serious problem for the individual: the problem of guaranteeing that other individuals will behave in ways that allow the individual to reach its goals. And a crucial component of the solution for this problem is to be able to predict the effects of one's behavior on other individuals and the effects of the behavior of other individuals for oneself. Hence, we should expect that human beings will be very preoccupied by the social causes and consequences of behavior, both one's own behavior and the behavior of others, and it is from this preoccupation that the discovery can emerge that other individuals also have a mind and that social interaction can be interactions between minds, i.e., intersubjectivity. Let us see how this can be so.

During their social interactions human beings discover that some inputs that arrive to their neural network also arrive to the neural network of other individuals since they are able to predict the behavior of other individuals on the basis of these inputs arriving to the other individuals' neural network. In contrast, there are other inputs that arrive to their neural network which are of a different type in that they are unable to predict the behavior of other individuals on the basis of these other inputs. If a strong light dazzles me and I see another individual that behaves as if he or she is also dazzled, I conclude that the light belongs to public reality. If I feel hungry but I see no other individual near to me that behaves as if he or she is also hungry or who says he or she is hungry, I conclude that the hunger I feel belongs to my private reality. Similarly, if I have in this moment a memory of something that happened to me in the past, I cannot predict nothing about the behavior of other individuals based on the fact that I have this memory, unless I communicate this memory of mine to another individual, thereby transforming this private input into a public input. In this way, the mind, which has begun to differentiate from my body (see above), begins also to differentiate from the mind of others and it becomes "my mind".

But the discovery of "my mind" is accompanied by the discovery that other people also have a mind. A possible way in which the mind of others emerges in the world of the individual is the following. As we have said, a crucial precondition for sophisticated social interactions is the cognitive capacity to learn to predict the consequences of the actions of others. (This does not necessarily mean that the cognitive capacity to predict has emerged evolutionarily or does emerge developmentally before the ability to interact socially in sophisticated ways. Cognitive complexity, including the capacity to predict the consequences of one's own actions and of the actions of others, may have emerged under the pressure to have sophisticated social interactions, not necessarily under the pressure to have sophisticated interactions with the physical world, for example to manipulate and change it. For discussions on the social origins of human intelligence, cf. Byrne and Whiten, 1988; Whiten and Byrne, 1997.) Now, when a human being is trying to predict the consequences of one's own behavior on the behavior of others and therefore to predict the behavior of others, he or she may discover that to hypothesize a mental life in the other individual, by interpreting him or her as possessing mental images, thoughts, rememberings, planned actions, and goals, results in more effective and correct predictions than interpreting him or her as simply capable of behaviors. Thus, "his or her mind" emerges together with "my mind". To postulate the existence of mental activities and mental contents in the other individual as factors that determine how the individual will behave and in particular how the individual will react to my behaviors, allows me to predict more effectively how the individual will behave and, therefore, it will make my behavior more effective insofar as my behavior has the goal to make the other individual behave in ways that benefit me.

As soon as the reality in which an individual lives includes not only his or her own mind but also the mind of others, the conditions for intersubjectivity are created, that is, for social interactions in which what interacts are minds, not bodies.

## 6. Two kinds of actions that can influence neural networks

A fundamental distinction in the study of behavior is the distinction between the cognitive and the dynamical aspects of behavior. Cognition is what an organism is able to do, the organism's capacities, knowledge, intellect. Dynamics is what drives the organism to do what the organism does, the organism's motivations, emotions, desires and fears. The hypothesis we want to propose to explain this distinction is that the distinction is linked to the existence of two different types of causes that can have effects on a neural network.

The first type of cause that can influence what happens inside a neural network is some event or process outside the neural network or even inside the neural network (cf. section on mental life) that causes a specific pattern of activation in some subset of the network's units. The activation pattern triggers a process of activation propagation which has specific properties determined by the network's specific architecture of connections (which unit is connected with which unit) and by the specific weights of the various connections. This is the type of influence on a neural network which is more often studied in the neural network literature. It is an influence that justifies speaking of neural networks as "information processing" systems or devices. "Information" is what is encoded in the initial activation pattern and this information is "processed" inside the network by transforming activation patterns into other activation patterns in the successive layers of units.

What are the characteristics of this first type of influence on a neural network and of its consequences, both immediate and long-term, for the functioning of the neural network? An activation pattern is something very specific which is observed in a very specific subset of the network's units and which is qualitatively, rather than quantitatively (intensively), distinct from other possible activation patterns. Furthermore, an activation pattern produces very specific effects inside the neural network in that the activation propagation triggered by the activation pattern goes through a specific architecture of connections inside the neural network, each with its own specific weight. Finally, the effects provoked inside the neural network by this first type of influence tend to leave specific and permanent traces in the neural network, in the sense that the future functioning of the neural network, the way in which the neural network will respond to future inputs, will be specifically and permanently influenced by the specific activation pattern.

However, in addition to this first type of influence on a neural network there exists a second type of action that can have an influence on a neural network and that can contribute to determining the network's functioning and, therefore, the organism's behavior. Chemical molecules diffused in the space among the network's units can influence the activity of the network's units and the way in which pre-synaptic units influence post-synaptic units, for example by changing the units' activation thresholds or by enhancing or depressing the activity of the neurotransmitters that mediate the communication between units.

What is the origin of these chemical molecules that can influence the functioning of a neural network? They can originate from inside the organism's body, they can be absorbed from outside the body (injected or ingested), or they can be produced as the output of specific units in the neural network itself. In all cases these molecules have a type of influence on the neural network which is very different from the influence that results from the first type of causes, i.e., activation patterns. The influence is now diffuse, that is, is the same type of influence acting on entire groups of units which are close together in space, is quantitative (intensive) rather than qualitative, and it does not

reflect necessarily the complicated network of connections and the finely adjusted pattern of connection weights of the neural network. One might characterize the first type of influence as topological while this second type of influence is topographic. The first type of influence is determined by the topological structure of the neural network, where what is critical is which unit is connected with which unit but not the physical (spatial) distance between units. The second type of influence is determined by the topographical i.e., spatial, structure of the neural network, where what is critical is the physical distance between units, not whether they are connected or unconnected. Furthermore, unlike the first type of influence, the effects of this second type of influence tend to be short-term effects, in that for example the modification that this second type of influence determines in the units' activation threshold tend to be temporary. In relatively short time the units' activation thresholds may return to the value they had previously. (For some simulations of diffuse and topographical processes influencing the functioning of neural networks, see Husbands, Smith, Jakobi, and O'Shea, 1998.)

While the first type of influence underlies the organism's capacities and cognitive activities, what the organism knows and knows how to do, the second type of influence determines the motivational, emotional, dynamical aspects of the organism's behavior, the different vigilance levels of the organism, its altered consciousness states, etc. An interesting hypothesis is that the two types of influences underly the two different but related ways in which one can try to help an individual with psychological problems: psychotherapy relies on the first type of influence, psychological drugs on the second.

## **7. Subjective experience**

Some of the processes or events that take place inside a person's nervous system occur with the person knowing nothing of these processes or events, while other processes and events results in subjective experiences for the individual. How can we account for this difference using neural networks?

Before we proceed we note that the presence/absence of subjective experience does coincide with the distinction between private and public inputs (see Section 3). An input is private if it is produced by causes which, for physical reasons, cannot produce the same input for the neural network of other individuals. However, an input can be private without being accompanied by subjective experience. My body sends lots of inputs and other influences to my brain and my brain responds to these private inputs and influences without my being aware of anything. On the other hand, subjective experiences always result from private inputs and therefore are always private. In other words, withing the general class of private inputs there are some that give rise to subjective experiences and others that do not. How can this happen?

An hypothesis is that subjective experience, i.e., the fact that some processes/events in the brain are felt, lived, experienced, whereas other processes/events are not, results from selective attention. In any given instant the nervous system of an organism receives a multitude of different inputs. Some of these inputs can be coordinated and integrated in such a way that the neural network takes into consideration all of them in generating an output. In other cases, however, this is not possible in that each input poses a different problem and requires a different output. In these circumstances, which are quite common, selective attention mechanisms enter the picture. Selective attention mechanisms are mechanisms inside the neural network that block the influence on the network's activity (activation propagation) of all of the current inputs except a single one, or a single set of coordinated inputs, and therefore cause that single input to determine the network's output. This selective process involves a sort of competition among the different inputs for the control of the network's output. The hypothesis, admittedly rather metaphorical at this stage, is that the inputs

which give rise to subjective experience are those that “speak louder” to make themselves heard and win the competition with other, concurrent, inputs (from outside of self-generated).

A second hypothesis one can advance to explain subjective experience is that subjective experience is linked to the self-generation of inputs which, as we have said in Section 4, results in mental life. In fact, mental life is necessarily associated with subjective experience. While the neural network can receive some inputs from outside and respond to these inputs without subjective experience, this seems to be impossible for mental life, i.e., for self-generated inputs (with the possible exclusion of actively “repressed” self-generated inputs). This may also help solve a problem that the philosophical notion of “qualia” can pose for our distinction between private and public reality. When I see the color blue of an object out there in the external environment I am having a public experience and I am living in a public reality since everyone nearby can see the same blue. However, as the philosophers have been ready to point out, there is a private side of this experience, a quality of “seeing the blue color” which does not appear to be public. How, then, can seeing the blue color be both private and public?

The activation pattern on the neural network’s input units which is observed when the individual sees the object with its blue color does not give rise, by itself alone, to any kind of experience. To have a subjective experience of seeing the blue color is something much more complex than just the occurrence of some particular activation pattern in the input units of one’s neural network. To have the subjective “qualia” experience of the color blue it is necessary for the neural network to propagate the input activation pattern in such a way that various self-generated inputs are produced by the network through its backward connections. These self-generated inputs which constitute the subjective “qualia” experience of that particular blue, are private.

## **8. The person and the photoelectrical cell**

What is the difference between a person who enters a room and sees that the room is lighted or dark, and a photoelectrical cell which is in two different states depending on whether the room is lighted or dark? Obviously, the seeing of the person and the “seeing” of the photoelectrical cell are not the same thing. But if we want to capture the difference between the two “seeings”, how useful is to say that the person is conscious of what he or she sees whereas the photoelectrical cell is not? Does this really explain anything? Our answer has been No. We have proposed a different route. What we should ask is: What should we add to the photoelectrical cell to be justified in saying that the entire thing sees in the same way as the person sees?

The photoelectrical cell resembles the neurons that make up the person’s retina, the visual receptors. In both cases the light energy present in the room produces an effect in the photoelectrical cell or in the person’s retinal receptors such that their state is different than the state which is observed when the room is dark. However, for the photoelectrical cell this is the end of the story. In the case of person, it is only the beginning. Beyond the retinal receptors there is an entire brain, an enormous network of cells with a very specific organization constituted by the architecture of connections and by the connection weights of the different connections. Unlike the photoelectrical cell, for the person the change of state in the visual receptors is nothing but the beginning of a chain of causes and effects that is propagated inside the person’s brain. In the brain the initial change of state in the visual receptor results in a succession of changes of state in successive groups of neurons, and some of these resulting states can be states of output neurons, that is, neurons that have effects outside the brain.

One of the internal states caused by the state of the retinal receptors may happen to be similar to the internal state caused in the past by another input in the retinal receptors or in the receptors of other

sensory modalities, for example, acoustical receptors. This may be favoured by the fact that the person's brain has a rich circuitry of backward connections. Hence, the person reacts to the light vs the dark with a mental image or a memory of something past, whereas this simply cannot happen in the case of the photoelectrical cell. Or, more complexly, the light or dark at the level of the retina may trigger an entire succession of self-generated inputs which in turn cause other self-generated inputs. This results in subjective experience for the person, which the brain-less photoelectrical cell will necessarily lack.

But behind the person's visual receptors there is not only a brain, there is also a body, the person's body, whereas the photoelectrical cell stands all alone, without a brain and without a body. As we have said, one of the effects of the propagation of activation in the person's brain is that some particular neurons change their state and these may be output neurons, that is, neurons which are not connected to other neurons but which, depending on the state in which they find themselves, produce particular effects outside the brain. Some of these effects can take place inside the person's body. Therefore, the consequence of light or darkness on the person's retinal receptors can be that the brain produces some specific changes inside the person's body. These changes in the body in turn may return to the brain, influencing the brain in the more diffused way we have discussed in the previous section (our second type of action that can influence a neural network). Thus, by seeing the light the person can feel relieved, or, by seeing the dark, he or she can feel fear.

But the fact that the person has a body whereas the photoelectrical cell has no body has other consequences that make the seeing of the person different from the "seeing" of the photoelectrical cell. The body, or its different parts, can be moved by the person's brain and, as we have seen, this will influence the further inputs that the person's brain will receive from the room. Seeing is not just receiving visual input from outside but it is noticing "sensory-motor contingencies" (O'Regan and Noe, in press), that is, noticing how visual inputs are influenced by one's motor outputs. The photoelectrical cell has no body, hence it cannot influence its visual inputs, hence it cannot see.

## References

Byrne, R. W. and Whiten, A. (eds.) Machiavellian Intelligence. Social Expertise and the Evolution of Intellect in Monkeys, Apes and Humans. Oxford, Clarendon Press, 1988.

Cecconi, F. and Parisi, D. Learning to predict the consequences of one's own actions. In R.Eckmiller, G.Hartmann, G.Hauske (eds.), Parallel Processing in Neural Systems and Computers. Amsterdam, Elsevier Science Publ.B.V. (North-Holland), 1990, 237-240.

Dehaene, S. The Cognitive Neuroscience of Consciousness. Cambridge, Mass., MIT Press, in press.

Edelman, G. M. and Tononi, G. A Universe of Consciousness. How Matter Becomes Imagination. New York, Basic Books, 2000.

Elman, J. L. Finding structure in time. Cognitive Science, 1990, 14, 179-211.

Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences USA, 1982, 79, 2554-2558.

Husbands, P., Smith, T., Jakobi, N. and O'Shea, M. Better living through chemistry: evolving GasNets for robot control. Connection Science, 1998, 10, 185-210.

Metzinger, T. Neural Correlates of Consciousness. Empirical and conceptual questions. Cambridge, Mass., MIT Press, 2000.

Nolfi, S., Elman, J.L. and Parisi, D. Learning and evolution in neural networks. Adaptive Behavior, 1994, 3, 5-28.

Nolfi, S. and Floreano, D. Evolutionary Robotics. Cambridge, Mass., MIT Press, 2000.

O'Regan, J.K. and Noe, A. A sensorymotor account of vision and visual consciousness. Behavioral and Brain Sciences, in press.

Parisi, D., Cecconi, F. and Nolfi, S. Econets: neural networks that learn in an environment. Network, 1990, 1, 149-168

Parisi, D. Neural networks and Artificial Life. In D. Amit and G. Parisi (Eds.) Frontiers of Life. San Diego, Academic Press, in press.

Rumelhart, D.E. and McClelland, J.L. Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Volume 1: Foundations. Cambridge, Mass, MIT Press, 1986.

Whiten, A. and Byrne, R. W. Machiavellian Intelligence II. Extensions and Evaluations. Cambridge, Cambridge University Press, 1997.