

Chapter 2

Emergence of Intentional Procedures in Self-Organizing Neural Networks

Henri Atlan and Yoram Louzoun

We have used a neural network formalism in order to analyze under which conditions a positive answer could be given to the following question: can neural networks self-organize so that not only structures and functions not explicitly programmed emerge from their dynamics, but also goals for intentional actions, set up and achieved by themselves?

Such mechanistic models of intentional self-organization are useful in that they allow to circumvent the usual circular explanation of intentionality by causal effects of assumed intentional mental states on bodily movements.

From a mathematical and modeling point of view, we have presented a simulation model for the analysis of intentionality through the study of intentional actions (Louzoun and Atlan 2007). We limit ourselves in this paper to the cognitive interpretation and the philosophical analysis of the obtained results. Intentionality in the psycholinguistic sense of “meaning” – where there is no “goal” except for the content of a thought in an internal deliberation of a sentence *meant* to say something – is left outside the scope of this work. We have limited ourselves to intentionality in a pragmatic sense as it is observed in intentional actions to solve two problems of causality: the apparent time inversion involved in final causes and the “*mind–body*” causal relationship involved in the usual picture of a mental state being the cause of bodily movements and actions.

The system we have developed is designed to devise new goals by itself and to reach these goals. The goals are determined by the capacity of a network to learn a relation between effects and the events that caused them. The model is a metaphor for the psychophysical goal learning process in cognitive beings. This process involves the ability to predict rapidly the result of a set of events, so that an initial

H. Atlan (✉)

Human Biology Research Center, Hadassah University Hospital, Ein Karem, Jerusalem, Israel
and

Ecole des Hautes Etudes en Sciences Sociales, 54 bd Raspail, 75270 Paris, France
e-mail: atlan@ehess.fr

Y. Louzoun

Math. Department, Bar Ilan University, Ramat Gan, Israel, 52900
e-mail: louzouy@math.biu.ac.il

event is reproduced knowing its expected result. In other words, prediction (which is knowledge) and intentional action are closely related. That is why this capacity is modeled using a non-supervised learning network associated to a recurrent neural network. However, while the prediction capacity is obviously based on memory of previous experience, this knowledge must be allowed some degrees of freedom, which produce new predictions of new events and the achievement of new goals. In our model, this capacity is simply the result of network dynamics where closely related but different states are associated in basins of attraction.

To summarize, the recurrent network represents a mechanistic causal process that develops from a random initial state to a steady state. There is no trivial relation between the steady state and the initial state of the network. The very indirect relation between initial and steady state represents the complexity of the causal relation in real environments.

The feed-forward learning network creates a link between final and initial states, allowing the time inversion occurring in goal-directed action. The input to the network is the steady state of the recurrent network, and the output of the feed-forward network is an initial state of the recurrent network, which is equivalent to a dynamic memory.

The selection mechanism chooses which final states are defined as goals, and works like a non-programmed satisfaction function, emerging from the partially random history of the system in its environment.

Our model is obviously not directly related to mental processes in its details. It only represents a plausibility analysis to show that the self emergence of meaningful *actions* is possible and can be explained by a relatively simple mechanism. The model allows us to study, which mechanisms are essential to have such a representation. The same question can also be addressed from the point of view of Spinozist monism as will be further discussed. The combination of a simple model and Spinoza's propositions enable us to provide plausible answers to shed new light on experimental results and propose ways to treat some of the most basic questions in cognition.

2.1 Minimal Necessary Requirements

Goals emerge in our simulation from a combination of four elements: A seemingly random process relating the initial and final states (which is actually a deterministic process too complex to be directly deduced from the initial and final states), a limited memory capable of remembering the relation between some initial and final states, a learning algorithm that invents a systematic relation between final and initial states, and an evolving set of required final states selected semi-randomly according to the frequencies of their appearance. Note that goals would not emerge in the absence of any of these elements. Thus, we think that our networks represent a minimal structure where such goals can be obtained. Obviously one can extensively alter the details and even completely replace the mechanistic aspect of each component. However the same general elements must prevail in order for goals to emerge.

- The first element required is an indirect dynamical link between initial and final states. Learning the relation between a state and its direct result is not defined as goal emergence. We define a learnt goal as a relation between an initial state and a final state that cannot be directly guessed from the initial state. Another aspect of the required dynamics is a difference between the probabilities to reach different final states. If all final states are reached with equal probabilities, the goal emerging would only be a mirror of the network history and would not represent an inherent property of the network.
- Memory is obviously required; it actually is the most important element of the network. The seemingly minor role of remembering a relation during the learning process is actually essential. In the absence of memory, the network would not be able to retrieve an initial state from a final state. This “time inversion” is the element giving the network a future prediction capacity. In other words, the network is able to predict the future in certain conditions, since it has seen similar evolutions and has learnt (either “erroneously” or “correctly”) a relation between an initial state and the final state it led to. A similar conclusion can be drawn for human behavior. Humans predict the future, since they have seen similar evolutions in the past and have learnt (either erroneously or correctly) a relation between a situation and its results.
- The learning algorithm is required since the capacity to attain goals depends on the ability to find a “simple” rule relating some of the initial states to the appropriate final states. Again one can infer from the network to human behavior, one can predict the future, only in cases similar to past events. These past events and their results were learnt and a time inversion mechanism is used to relate new situations to their future.
- Finally, the evolving set of goals allowing for both stability and newness is required in order to distinguish between goals that can and cannot be learnt, and goals for which no simple rule can be obtained. A specific aspect of the goals that we have requested in the current application is stability (i.e. we required that goals should change slowly compared to the network dynamics). This request is not essential. One could imagine rapidly changing goals (e.g. the mind of a small child). However, most aspects of human behavior are based on a set of relatively long term goals. The emergence of these “long term goals” is equivalent to the emergence of stable goals in the current application. This element is thus not required for the goal emergence per-se, but it adds an aspect of *meaning* to the goals. In addition, the possibility of newness is embedded into the role of small random variations in the definition of goals.

2.2 Externally Versus Internally Defined Goals

In the current application, we minimized the model and merged together two different tasks. The memory device which allows for goal directed action and the learning device which allows for goal definition by the system itself are merged into the operation of the feed-forward, perceptron-like, network.

Of course, the two different tasks performed by the feed-forward network can be separated, especially if the model is designed in a more trivial fashion to achieve predefined goals, assigned from outside the system. Contrary to a goal defined from outside, a self-generated, internally defined goal is not a goal because it has some inherent value from the beginning. It is a goal because it represents a properly learnt and stable link between initial and final states. A set of such goals is an emerging and stable property of the network's structure and the history it underwent. Dependence on history represents how the system adapts itself and generates new goals accordingly. On the other hand, externally predefined goals can be learnt more simply. Each of them must be coded into an attractor state of the recurrent network; and then kept in memory as one of the final states to be eventually retrieved with an initial state leading to it from its basin of attraction. (It is clear that only attractor states of the recurrent network can be established as goals, either by external imposition or by non-supervised learning, since a state can be stored as a goal only if the system can reach it with a high enough probability.)

2.2.1 *At the Beginning*

For example, one could consider that the initiation of the learning process needs not start from scratch, as in the present model. Before learning, a basic set of goals may have been stored as an initial set of "instinctual" goals with which to start. This may be the result, in the real world, of *long term evolutionary processes*, which may be simulated, for example, by genetic algorithms driven by selection for survival. Such processes must be distinguished from the mechanisms of setting oneself cognitive goals that is studied in the present work. Such a priori goals may produce built in, basic drives to start with, like biochemical signals for hunger, sexual drive, tissue damage repair and so on. These signals would affect only the initial set up of goals, but not the general mechanism of goal development. One can even set a permanent "vital" set of goals selected through a long evolutionary process. These goals can be hard wired not to change. Another possibility would be that some goals have an inherent higher score than others. We have tested models to include such initial goals or preferred goals, and the subsequent picture emerging from these models is similar to what we currently report.

According to our model, intention and action appear to be one and the same realization, simply represented in different ways. This implies that an intention to act is always normally associated with its execution. In other words, both the action and the intention are represented by links between initial and final states. The difference between the action and the intention is actually the difference between an action actually performed and its initiation, as indicated by neurophysiological data discussed further. This difference results in our capacity to stop an action once initiated. We would call an action interrupted after being initiated, an intention to do an action and invent a mental state to represent it. This view is opposed to the usual mentalist assumption that an intention exists first in the mind as a "pure" mental

state, possibly, but not normally associated to its execution. In our model, as in the work of Anscombe (1957), intentions are not defined as pure subjective states of the mind, but as properties of some sets of actions, which make them intentional and different from non-intentional ones. The fact that a subjective intention to act may not be followed by its execution is not to be seen as the normal flow. Rather, it must be related to an external obstacle to the execution or to any other kind of superimposed inhibition preventing the iterative process to reach completion.

One does not need to invent intentional mental states as causes of teleological actions. This is of course in contrast with common sense or folk psychology based on our initial insight of the causal relation between will and action. However, neurophysiology data on voluntary movements contradict this commonly accepted picture as well and support our model in showing that the conscious will to trigger an action does not necessarily precede the action.

2.3 Neurophysiology of Voluntary Movements

Following observations by Benjamin Libet and his co-workers (Libet et al. 1983; Libet 1985, 1992), recently confirmed and expanded (Haggard and Eimer 1999; Haggard et al. 2002), spontaneous short-term conscious decision to act with no pre-planning does not precede but follows by approximately 300 ms the *initiation* of movement, as measured by the Readiness Potential cortical activity. Thus, initiation of a voluntary action is triggered by some unconscious activity, and the following awareness is interpreted as its cause. When asked about the timing of their decision, subjects perceive it, by antedating, before the initiation of the action. However, the motor activity itself follows by 150–200 ms the conscious decision to act, which means that a conscious “veto” is possible, as an inhibition of the movement after its inhibition.

Most of the controversy around this work was triggered by the difficulties to reconcile these data with the traditional Cartesian concept of free will and to integrate these data within the commonly accepted mentalist causal theories of action. The model presented in our work contributes to make these data intelligible within an alternative monist theory of action. Mentalist theories of action, based on the idea that mental representations described as subjective states of the mind, can cause objective brain states able to trigger physical movements, were extensively analyzed and criticized already in 1957 in a philosophical and psychological context (Anscombe 1957). This criticism, as well as our model, contradicts our common sense representation of free will as a direct cause of voluntary actions. However, the general question of free will as an illusion or a reality remains open, because the model, as do Libet’s data, allows believers in free will to relate it in an indirect way, to a possibility of vetoing a movement after its initiation, rather than to the initiation itself.

Antedating the conscious decision to act may be thought of as a temporal illusion (analogous to a spatial visual illusion), with a possible adaptive value whereby

voluntary actions are linked to our memory-based capacity of prediction and self-awareness (e.g. (Llinas 2001)). As in our model, inhibition of movement completion after initiation explains intentional action with no execution. However, this does not necessarily infer that the problem of free will is solved in one way or another: if one can relate it to vetoing the execution of a movement, one cannot exclude, on the other hand, that vetoing itself would be caused by a non-conscious event, in spite of our spontaneous subjective conscious experience.

Thus, the question of whether free will is an illusion or not is definitely left outside the scope of this study. Similarly, long-term deliberation leading to intentions to do something in principle with no specific timing for the actual decision to act, are left outside Libet's observation. In the experimental setting, the patients were asked to perform some movement and to decide upon the timing. It is clear that their very participation in the experiment indicates their agreement and intention to do it before their decision.

2.4 Philosophical Interpretation

One feature of the views presented here is the monist ontology involved in the approach to the mind–body problem. Spinozist philosophy is certainly the most radical monist attitude towards this problem. This is apparent, for example, in propositions such as

“Body cannot determine mind to think, neither can mind determine body to motion or rest or any state different from these, if such there be” (The Ethics, III, 2), where Spinoza denies the possibility of causal relationships between the mind and the body, not because they would pertain to two different substances, as in Descartes, but precisely because they are *“one and the same thing, though expressed in two ways”* (Ibid. II, 7, note).

The analysis of some aspects of this psycho-physical monism will help to better understand the philosophical counterintuitive implications of our model, as well as of the neurophysiological data on voluntary movements briefly reported in the previous section.

Let us first note that this Spinozist denial of a causal relationship between mind and body states, just mentioned, implies that the cause of a voluntary bodily movement must always be some previous bodily (brain) event or set of events, and not a conscious decision viewed as a mental event as described by subjective reports about conscious experiences. The difference from a non-voluntary movement is the nature and degree of conscious experience accompanying it. But in any case, a conscious mental event in this context may accompany the brain event *but not be its cause, being in fact identical with it*, although not describable in the same language. Results from neurophysiology support this view: unconscious initiation of voluntary action precedes the conscious decision to trigger the movement. Thus, our model may provide Spinozist monism, however counterintuitive, with some theoretical and philosophical interpretation.

This kind of counterintuitive identity between different properties or events, identical but not describable by synonymous enunciations, was called a “synthetic

identity of properties” (Putnam 1981), to be distinguished from the usual analytical identity, where synonymous descriptions can replace one another. Hilary Putnam found an example of synthetic identity in the notion of physical magnitudes, which we employ in physics, such as “temperature” and “mean molecular kinetic energy” being synthetically, but not analytically, identical. In the same context, Putnam explicitly related the Spinozist psycho-physical identity to such a synthetic identity, as a way to overcome many well known difficulties in understanding this approach to the mind–body problem (see also (Atlan 1998a)).

Similar results on affects and emotions, indicating a lack of causality between body and mind, have been proposed by A. Damasio, with the same reference to Spinozist monism as its philosophical interpretation (Damasio 2003).

This stance, as well as the elaborated Wittgensteinian view of intentional descriptions (Wittgenstein 1953), has been neglected by most philosophers and cognitive scientists, mostly because it contradicts our common sense experiences and the commonly accepted ethical implications of free will which go with them. Thus, under the influence of mentalist theories in psychology (for analysis and criticism see e.g. Anscombe 1957; Davidson 1970; Fodor 1981; Shanon 1993; Chalmers 1995), intentions are viewed as some kinds of conscious mental states, able to cause bodily movements whenever an intentional action is executed. These theories raise several difficult questions, such as:

1. How can a mental state be the cause of a physical movement?
2. More generally, what is the conscious intentional experience made of?

The first question has been addressed, more or less successfully, by several philosophers. Among them, Donald Davidson’s theory of action may be the most comprehensive (Davidson 1970, 1999), especially in view of his definite monist attitude, which he explicitly relates to *The Ethics* of Spinoza. However, his willingness to stick to common sense conscious subjective and ethical experiences does not allow him to overcome serious difficulties in trying to reconcile the Spinozist explicit denial of causal relationship between subjective states of mind as such and objective bodily movements, with his “anomalous monism” (Davidson 1991; Atlan 1998a).

The second question covers several problems related with different aspects of what we call consciousness. According to David Chalmers (1995), some of these problems are “easy”, although not trivial: they deal with specific cognitive aspects of consciousness, related with objective mechanisms accounting for cognitive properties, such as memory, learning, adaptation, etc. However, what he calls the “hard problem” is the “question of how physical processes in the brain give rise to subjective experience”. This question is the same in the opposite direction as that of intentional actions, where subjective intentions are supposed to cause physical movements.

In our work, we depart from mentalist causal theories of action and we try to come back to a more objective approach to the question of causality (Atlan 1998b). The model presented here exhibits one of the main features outlined by Anscombe in order to circumscribe the logical difficulties of these theories, namely the approach of intentionality through the study of intentional *actions*. As mentioned above, this implies that intentions and actions are not dissociated to start with, and that the

normal state of affairs is the execution of the intention. Such a dissociation, which may occur when an intention is not accompanied by an action, is the result of an obstacle or inhibition of the execution.

In this view, the “hard problem” of causality between the mental and the physical is eliminated: there is no causal relationship between an intention as a mental state and action as a bodily movement, because “roughly speaking, a man intends to do what he does” (Anscombe 1957). Because this view seems counter-intuitive and raises new questions, following the quest initiated by Wittgenstein about the status of intentional statements language games, Anscombe feels compelled to add: “But of course that is *very* roughly speaking. It is right to formulate it, however, as an antidote against the absurd thesis which is sometimes maintained: that a man’s intended action is only described by describing his *objective*”. In many instances the objective of the agent is a description after the fact, aiming at answering the question: “Why did you do it?”.

Let us conclude with several features of the non-mentalist model of intentions presented in this work, which appear almost literally in Spinoza’s writings, at the point that one could speak of a “Spinozist neurophysiology”.

1. Decision to act and previous knowledge allowing prediction are two different aspects of the same process associated with voluntary actions, although the former seems directed towards the future and the latter towards the past. That is the case because intentions are described by means of intentional *actions* and not of intentional mental states as causes of the actions. “*Will and understanding are one and the same*” ((Spinoza 1677), II, 49, corollary) seems to be an abrupt statement of this counterintuitive concept.
2. In our model, general sets of goals are memorized from learning by experience. The acquired knowledge results from the interaction between the internal structure of the network and the history of its most frequent encounters with classes of stimuli from its environment.

In the context of the classical controversy about the reality of “Universals”, we read:

... these general notions (called Universals) are not formed by all men in the same way, but vary in each individual according as the point varies, whereby the body has been most frequently affected and which the mind most easily imagines or remembers. For instance, those who have most often regarded with admiration the stature of man, will by the name of man understand an animal of erect stature; those who have been accustomed to regard some other attribute, will form a different general image of man, for instance, that man is a laughing animal, a two-footed animal without feathers, a rational animal, and thus, in other cases, everyone will form general images of things according to the habit (disposition) of his body ((Spinoza 1677), II, 40, note).

Thus, this “disposition of the body” is made by the way the cognitive system (mind–body) is assembled and also by the way it has been most frequently affected.

3. According to the neurophysiological data on voluntary movements reported before, as well as in our model, voluntary action is triggered by some unconscious stimulus, accompanied but not caused by a conscious state of the mind. A conscious observation with an understanding of our action accompanies that action

but is not its cause. And we can interpret it as a decision of our will which determines the action, because we do not know the unconscious events in our body which are the real causes.

Now all these things clearly show that the decision of the mind and the desire or decision of the body are simultaneous in nature, or rather one and the same thing, which when considered under the attribute of Thought and explained through the same we call a decision, and when considered under the attribute of Extension and deduced from the laws of motion and rest we call determination ((Spinoza 1677), III, 2, note).

4. As noted in Libet's observations there is a slight delay between the triggering of action and our being conscious of it, because consciousness and understanding take time: as in our model, they need to be retrieved from *memory*. In other words,

we can do nothing by a decision of the mind unless we recollect having done so before ((Spinoza 1677), III, 2, note).

5. In the stance adopted here, we obviously *lose* something, namely common sense about free will and causation of actions by decisions of a non-bodily mind. However, we *gain* understanding of intentional actions without resorting to hidden causal properties of mental states. Let us note that the reality of free will is not necessarily denied, although its content is modified. According to Libet, it can be located in a kind of *veto* function, i.e. a possible inhibition of movement after it has been initiated. In addition, nothing is said here about the possible effects of long term deliberations and decisions to act "in principle", with a more or less extended period of time until the decision is made to start the action. Spinoza's stance about free will is more radical:

...men think themselves free on account of this alone, that they are conscious of their actions and ignorant of the causes of them; and, moreover, that the decisions of the mind are nothing save their desires, which are accordingly various according to various dispositions of their and other interacting bodies ((Spinoza 1677), note on proposition III, 2, mentioned above).

6. At last, the picture of intentional actions presented in this work helps to better understand what "desire" in the practical syllogism is about¹: an unconscious drive with awareness of the goal which one is driven to.

¹ Let us recall the classical description of intentional actions by the practical syllogism:

- Agent A desires to be in state S.
- A knows or believe that C is a cause for S.
- Therefore A performs C.

This description assumes intentional mental states from the beginning, such as desire, knowledge, belief. In our model, knowledge or belief are just retrieved memories of previous causal events. In addition, as Elizabeth Anscombe rightfully noticed, the first proposition of the syllogism may be conflated with the third. Contrary to the usual demonstrative syllogism (Men are mortal, Socrates is a man, etc.), the first proposition here does not add information: it is contained in the "therefore" of the third proposition. Our model may be seen as a computer simulation of this modified syllogism, where intentional mental states causing intentional actions and different from them are not needed.

This definition of desire has been extended further by Spinoza to the realm of moral judgements:

Desire is appetite with consciousness thereof. It is thus plain from what has been said, that in nocase do we strive for, wish for, long for, or desire anything, because we deem it to be good, but on the other hand we deem a thing to be good, because we strive for it, wish for it, long for it, or desire it ((Spinoza 1677), III, 9, note).

References

- Anscombe GEM (1957) *Intention*. Basic Blackwell, London
- Atlan H (1998a) Immanent causality: a spinozist viewpoint on evolution and theory of action. In: Vijver Gvd (ed) *Evolutionary systems*. Kluwer, The Netherlands, pp 215–231
- Atlan H (1998b) *Intentional self-organization. Emergence and reduction. Towards a physical theory of intentionality*, Thesis Eleven 52:5–34
- Chalmers DJ (1995) The puzzle of conscious experience. *Scientific Am* 12:62–68
- Damasio A (2003) *Looking for Spinoza. Joy, sorrow, and the feeling brain*. Harvest Books, Harvest edition, Washington
- Davidson D (1970) Mental events experience and theory. In: Foster L, Swanson J (eds) *University of Massachusetts*, Amherst, pp 79–81; reprinted in Davidson D (1980), *Essays on actions and events*. Oxford University Press, New York
- Davidson D (1991, 1999). Spinoza's causal theory of the affects. In: Yovel Y (ed) *Ethica III. Desire and affect. Spinoza as psychologist*. Little Room, New York, pp 95–111
- Fodor JA (1981) *Representations*. MIT, Cambridge, MA
- Haggard P, Eimer M (1999) On the relation between brain potentials and the awareness of voluntary movements. *Exp Brain Res* 126:128–133
- Haggard P, Clark S, Kalogeras J (2002) Voluntary action and conscious awareness. *Nat Neurosci* 5(4):382–385
- Libet B (1985) Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behav Brain Sci* 8:529–566
- Libet B (1992) Models of conscious timing and the experimental evidence (Commentary/Dennett and Kinsbourne: Time and the Observer). *Behav Brain Sci* 15:213–215
- Libet B, Gleason CA, Wright EW, Pearl DK (1983) Time of conscious intention to act in relation to onset of cerebral activity (readiness potential): the unconscious initiation of a freely voluntary act. *Brain* 106:623–642
- Llinas RR (2001) *I of the vortex: from neurons to self*. MIT, Cambridge, MA
- Louzoun Y, Atlan H (2007) The emergence of goals in a self-organizing network: a non-mentalist model of intentional actions. *Neural Networks* 20:156–171
- Putnam H (1981) *Reason, truth and history*. Cambridge University Press, Cambridge
- Shanon B (1993) *The representational and the presentational*. Harvester Wheatsheaf, Simon and Schuster, New York
- Spinoza B (1677) *The Ethics* (engl. translation H.M. Elwes 1955). Dover, New York



<http://www.springer.com/978-90-481-3528-8>

Causality, Meaningful Complexity and Embodied
Cognition

Carsetti, A. (Ed.)

2010, XLVIII, 360 p., Hardcover

ISBN: 978-90-481-3528-8