Towards an anticipation control theory of mind

Francis Heylighen
Evolution, Complexity and Cognition group,
Vrije Universiteit Brussel

ABSTRACT: xxxxxxx

Introduction

The functioning of the mind or brain remains one of the great scientific mysteries. A few decades ago, the new discipline of *cognitive science* seemed to offer a general paradigm [De Mey, 19] that could tackle such abstract problems as thought, memory, perception and even emotion, by modelling the mind as an information-processing system.

The general idea is that information enters the brain through the senses, is then interpreted by the knowledge stored there, and exits in the form of decisions as to which action to perform, or which answer to produce to a given question or problem. This information is typically formalized as a string of logical or mathematical symbols representing the situation that is perceived. Knowledge is conceived as a system of rules or procedures for manipulating those strings. Different manipulations are tried out until the resulting strings have the characteristics desired for a solution to the given problem.

This modelling paradigm was at the basis of classical artificial intelligence (AI), allowing it to simulate problem-solving activities such as chess playing, manipulation of elements by a robot in a simplified environment ("blocks worlds"), or the way an expert makes a medical diagnosis. In spite of the great initial expectations, however, this program has basically failed to replicate human intelligence.

This has spurred the development of a number of complementary paradigms, the most successful of which are neural networks [], situated and embodied cognition [], and dynamical systems models [Beer; Port & Van Gelder; Thelen & Smith]. Each of these has met with partial successes, explaining or simulating phenomena that other paradigms had difficulty dealing with. Neural networks have done away with the requirement of discrete, symbolic representation and rule-based manipulation, while maintaining the linear,

input-output view of processing. Situated and embodied cognition has emphasized the on-going interaction or feedback cycles between cognitive system and environment. Dynamical systems theory has added the dimension of time, studying how a cognitive process unfolds continuously under the influence of internal and environmental "forces".

But an overarching theory of mind still seems far off. This has led to a more radical departure from cognitive science, attempting to develop a true *science of consciousness*. Much of the motivation comes from the apparent incapability of formal models to explain subjective experience, i.e. the intuitive "feel" or "understanding" that we have of the phenomena we perceive or conceive.

This problem was formulated perhaps most forcefully by Searle [] in his thought experiment of the "Chinese room". The person sitting in that room does not know Chinese, but gets handed bits of paper with strings of Chinese characters. Applying a complex system of written procedures, the person responds to each specific string by assembling and returning another specific string. If the procedures are sufficiently detailed to produce a plausible response to all the Chinese sentences that a Chinese speaker might like to submit, it would appear to that outside observer that something in the room "understands" Chinese, although the person sitting there merely mechanically assembles bits of paper without any grasp of what is going on.

AI researchers have correctly pointed out that although neither the person, nor the written procedures may have any understanding of the Chinese conversation that is taking place, the room as a whole, i.e. the emergent system formed by person, procedures, and papers, may be said to exhibit understanding. But neither AI supporters nor their critics have paid much attention to the absolute implausibility of a system of fixed procedures, whether written on paper or in computer code, that would be able to perform an intelligent, unconstrained conversation with a real human being. Such a conversation is fully context-dependent, which means that the answer to any given question will change with the most subtle changes in the situation, as the conversants "get to know" each other, taking into account previous exchanges, the conversants' moods, outside events, shared experiences, etc. Formal rules are much too rigid to steer such a continuous adaptation to the "feel" of the context.

The most radical departures from the cognitive paradigm have focused on the so-called "hard problem" of consciousness [Chalmers], noting that even if we would be able to build a robot exactly replicating the way human beings interact with the world, this robot or "zombie" would still lack the elusive quality of experience or "feeling" that accompanies such interaction. This leads to metaphysical speculations that subjective experience is some mysterious substance independent of any structure or function of the cognitive system as we know it, which perhaps is to be found in the paradoxical phenomena of

quantum mechanics that govern the interactions between atoms and molecules at the lowest level of our brain processes [Penrose].

The present paper wishes to propose a much more pragmatic approach to the problem of feel, experience or understanding. This approach encompasses the recent connectionist, situated and dynamical extensions to cognitive science, but integrates and extends them further, producing a more coherent and complete picture of the basic mechanism of mind. This perspective, which we will call the anticipation-control theory of mind, is inspired more by neurophysiology and the psychology of perception and attention than by the linguistic, logical and computational models that inspired AI. Yet, it does not require a detailed understanding of the brain's anatomy or physiology, merely a more realistic appraisal of what the brain actually does.

Recently, a number of authors [McCrone, Hawkins, Richardson, O'Regan & Noë] have brought forward different aspects and applications of this emerging new paradigm. But its roots can be traced back much further, to the ideas of the psychologists Hebb [194] and Neisser [197], and the general paradigm of cybernetics that emerged in the 1940's but which was largely eclipsed by AI in the 1960's. My personal contribution in this paper is to review and attempt to synthesize these different proposals, using the more recent paradigm of evolutionary cybernetics as an integrating theoretical framework.

The new framework in a nutshell

The new approach we are trying to define was probably described in most detail by Hawkins [2005], who calls it the *memory-prediction framework* for modelling intelligence, and by McCrone [1999], who calls it the *dynamicist* or *complex adaptive systems perspective* on the brain. Given that "dynamicist" is a rather vague term, which may moreover invite confusion with the related but more vaguely defined "dynamical" approach, whereas "memory" appears to focus on permanently stored information rather than on the dynamic processes that both authors emphasize, I propose to call the new perspective *anticipation-control.*

This immediately brings us to the core of the theory: a variety of psychological and neurological observations seem to indicate that a basic activity of the brain is prediction or anticipation of perceptions. Whenever the brain receives some stimuli similar to stimuli it has experienced before, it will use its stored experience to "fill in" or anticipate the further stimuli that are likely to follow. Thus, when from the corner of my eye I see a leg entering my field of vision, I will immediately anticipate the appearance of another leg, a body, two arms and a head. In the unlikely case that those expected stimuli would not appear and I would merely see an isolated leg dangling in front of

me, my attention would be grabbed and my whole brain would be aroused with activity trying to make sense of this incongruous phenomenon.

The principle is that the brain is extremely good at learning recurrent regularities—or more precisely *co-occurrences* i.e. elements or patterns that tend to appear together or in close succession. Since the world is full of such regularities (e.g. practically all human bodies have two legs, two arms and one head), this makes the brain very effective at inferring as yet unseen shapes, aspects, properties and consequences from the very incomplete and noisy information that it receives from the senses. It is the whole of these implicit anticipations induced by a phenomenon that, according to Hawkins [], determine our "*understanding*" of the phenomenon. This moreover allows the brain to maintain an *invariant representation* of the phenomenon even when the sensory input is fragmentary and constantly changing, as when the eye quickly scans different aspects of a scene.

It is only when anticipation obviously fails, as in the situation with the single, dangling leg, that we have to admit that we do not understand. In this case, the brain will try to correct its predictions, by searching for alternative regularities or additional information that may help it to make sense of the incongruous phenomenon. The view of a single leg may remind you of the plastic legs that are sometimes used in shop windows to exhibit stockings, while the additional information of a wire attached to the leg may direct your attention to the fact that some children are manipulating the leg from a distance. This, together with your knowledge of childrens' pastimes, brings a new understanding of the situation: the children are playing a game with a plastic leg in order to startle passers-by...

This illustrates the second main component of the theory: *control*. For the brain, any failure of anticipation constitutes an error that needs to be corrected. This entails a *feedback* process, from the cognitive function that noticed the anomaly back to the perception that triggered it. A novel aspect of the anticipation-control perspective is the observation that this feedback occurs on many different levels. As the situated and embodied perspective has argued, cognition heavily relies on sensory-motor feedback: the anomaly triggers a number of actions of the muscles, such as a redirection of your gaze or extension of your hands towards the anomalous object, in order to obtain additional information or test assumptions about the object. This on-going interaction between mind and world has been summarized by the slogan that *the environment is its own best model*. In other words, if your stored knowledge is insufficient to explain or predict a perceived phenomenon, check the phenomenon itself rather than your model of it.

But regulatory feedback also takes place *within* the brain, as many neurological experiments have demonstrated. The low level, default expectation that a leg is necessarily part of a body, when falsified will activate a higher level process looking for less common possibilities. If this higher level

process comes up with a plausible explanation, such as a mannikin's leg as seen in shop windows, it will feed back this interpretation to the lower level perceptual circuits, priming them to pay attention to specific symptoms or tell-tale signs associated with this new interpretation—such as perhaps a stocking knotted around the leg's top ending. If these signs can be recognized in the lower-level perceptual representation, the percept will undergo a Gestalt switch to a wholly different understanding, and aspects that were not noticed before may now become so obvious that you wonder how you could ever make the mistake to think that this was a real leg...

A final type of feedback forms the main focus of the connectionist perspective on cognition. A neural circuit that succesfully predicted what would happen gets reinforced; one that made a wrong prediction is weakened. In this way every experience of trying to anticipate phenomena leaves its trace in the organization of the brain, making it ever more effective at further anticipation. This is how we implicitly learn our knowledge.

A core tenet of the anticipation-control theory, as emphasized by Hawkins [2005], is that the brain is organized *hierarchically*, in different levels of invariance or control, whereby a higher level only becomes active when the lower levels fail to make a good prediction. It is this hierarchical structure with massive feedback from the higher to the lower levels, allowing us to recognize and learn ever more complex invariances and covariances of patterns, that distinguishes the anticipation-control model from traditional neural network models. These models consist of a mere two or three layers of artificial "neurons", connected in a feedforward network—i.e. without feedback. While such neural network models are good at learning to recognize basic patterns—such as the shapes of letters—even on the basis of noisy and incomplete information, they break down when these patterns appear in anomalous orientations or configurations, because they recognize only surface regularities, not higher level invariances.

The anticipation-control theory not only helps us to understand cognition and perception, but also the vaguely defined processes that are usually referred to as "consciousness". One way to clarify this complex and confuse concept, is to divide it into two basic aspects [Block, 1995]: access consciousness and phenomenal consciousness .

Access consciousness refers to the control we have over our cognitive processes, i.e. to our capability to explicitly monitor, examine, and manipulate our thoughts and perceptions. In the anticipation-control approach, access consciousness only appears at the highest level of the hierarchy, which handles the most complex and abstract cases of anticipation. It requires a high level of neural activation or arousal directed at the difficult problem, trying to systematically explore different invariant patterns that could be used to explain or tackle it. Most of our brain activity, on the other hand, is automatic or implicit, without conscious thought. The reason is that our brain is generally so

good at using and learning regularities that most of our perception and behavior can rely on our implicit anticipations, without any need to investigate or control our subconscious assumptions.

Phenomenal consciousness refers to what we have called subjective experience, "feel" or "understanding" of a phenomenon. O'Regan et al. [] have suggested that this feel (e.g. of driving a Porsche car) depends on the anticipations created by our implicit schemes for sensory-motor feedback, which remain active even when sensory or motor activity are temporarily suspended (e.g when with our hands unmoving on the steering wheel we temporarily close our eyes). However, this invites the question of how abstract categories, such as "redness" or "freedom" can produce specific feels (what philosophers call "qualia"). But in the anticipation-control theory even abstract categories correspond to higher level neural invariance structures, which trigger expectations of, and are triggered by, a whole range of other categories. Thus, considering the quality of "redness" may make you think of blood, warmth, fire, romance, danger, roses, a political party, etc. It is the whole of these—stronger or weaker, explicit or implicit—anticipations, which we learn through personal experience and which thus are different for every subject, that together can be said to constitute our "feel" or "understanding" of what redness means.

After this quick overview of the new theory, we are ready to delve deeper into its conceptual underpinnings and from there move to explanations of concrete, empirical observations.

Evolutionary-cybernetic foundations of cognition

Control

According to the perspective of evolutionary cybernetics, all living systems are goal-directed, with evolutionary *fitness*, i.e. survival and reproduction, as their overall, implicit goal. The better an organism is at achieving this goal, the more offspring it will leave, and therefore the more numerous this type of organism will become. Thus, given variation and natural selection, over the course of evolution organisms will become increasingly more effective at reaching this basic goal. As organisms become more complex, this abstract, implicit goal of fitness will be realized as a system of more concrete, explicit, subsidiary goals, such as: eat sufficient food, avoid danger, avoid cold, have sex, etc.

Cybernetics [Heylighen & Joslyn, Ashby,] has shown that goal-directedness is intrinsically a problem of *control*: to reach its goal, an agent needs to counteract or compensate any perturbation or obstacle that keeps it away from this preferred state-of-affairs. This implies a negative feedback loop, which reduces and eventually eliminates any deviation from the goal state.

Control moreover implies a rudimentary form of *cognition*, as the system must "know" which action to perform in order to compensate which perturbation. This knowledge can be expressed in the form of a set of condition-action rules:

if condition A occurs, *then* perform action B: $A \rightarrow B$.

As Conant and Ashby [] have shown, every good controller must incorporate a rudimentary model of the system being controlled, where a model is defined as a mapping from conditions or states of the system to regulatory actions. Therefore, as also noted by Maturana and Varela [], all living systems, being goal-directed, must also be cognitive. Even the simplest bacterium when confronted with a problem (e.g. the presence of a toxic chemical A), if it is to survive must "know" how to tackle this problem (e.g. by manufacturing the right enzyme B that can neutralize toxins of type A).

Anticipation

Negative feedback as a control mechanism has an essential shortcoming: to decide about the right counteraction to compensate a given perturbation, the perturbation must already have occurred. Feedback control is also called "error-controlled regulation", since it first must allow the deviation or "error" to become large enough to be detected before it can take counteraction. But there will be cases in which there is not enough time to take counteraction, either because the error erupts so suddenly that it destroys the agent before it can react, or because the agent simply needs a lot of time to identify and solve the problem. For example, if you suddenly slip off a steep cliff, it is too late to start climbing up again: you are already falling to your death hundreds of meters below. To avoid this, it is wise to look out for cliffs ahead and keep a safe distance away from their borders. In other words, you should try to anticipate the danger on the basis of any available signs, and take counteraction *before* the problem has occurred.

This defines the second major control mechanism: *feedforward*. Feedforward requires a more advanced type of knowledge, that maps present conditions not immediately to control actions, but to further, anticipated conditions. This can be represented as a *condition-condition rule*:

if condition **A** occurs, then expect condition **B**: $\mathbf{A} \rightarrow \mathbf{B}$.

The arrow $\rightarrow$ does not necessarily mean that B follows A, or is caused by A, as in the rule: **flame in gas-filled room** $\rightarrow$ **explosion**. It can also denote a logical implication, attribute, category membership or even a mere association, as in **banana** $\rightarrow$ **fruit**, **banana** $\rightarrow$ **sweet, banana** $\rightarrow$ **yellow,** or **banana** $\rightarrow$ **monkey**. It just

means that if something is recognized as a banana, then it can be expected to be a fruit, to be sweet and probably yellow, and perhaps even to be eaten by a monkey. These expected conditions can be mapped recursively to further expected conditions, and eventually to actions that tackle the present or anticipated problem, e.g. **do not light a match**, or **eat the banana.**

Uncertainty

Feedforward control or anticipation-based regulation too suffers from an essential shortcoming: since you never have perfect information and knowledge, you can never accurately predict everything that will happen. There will always remain uncertainty, and thus a potential for error. Moreover, as anticipations are extrapolated further into the future, errors, however small initially, will accumulate and grow, until there is hardly any relation at all between prediction and reality. That is why tomorrow's weather forecast is generally reliable, but a forecast for next month or next year is basically useless.

At best, we can estimate the probability P(s) that a particular state of affairs s will occur. Given such a probability distribution P, we can define the uncertainty or entropy H of our prediction, using the classic formula of Shannon:

$$H(P) = -\sum_{s \in S} P(s).\log P(s)$$

In practice, we do not know the absolute probabilities of different future conditions A. At best, experience may have given us an estimate of the conditional probability P(B | A) of B occurring given that A has been perceived. This can also be seen as a transition probability or expectation strength of a condition-condition rule: P(A → B). For example, P(**banana** → **fruit**) = 1 (i.e. the rule is logically necessary), P(**banana** → **yellow) =** 0.7 (likely but not certain), P(**banana** → **monkey**)= 0.05 (unlikely, but the conditional probability of seeing a monkey, given a banana, is still higher than the absolute probability of seeing a monkey).

The whole of these probabilities defines a transition matrix, which represents predictive knowledge as a Markov process. We can then define the conditional uncertainty H(P(B | A)) in the same way as above but now using the conditional or transition probabilities. This uncertainty measures the reliability of the agent's knowledge. The higher the conditional uncertainty, the less accurate the predictions B that the agent can make on the basis of the present situation A. Uncertainty or entropy will typically increase the further you extrapolate into the future, e.g. by calculating the probabilities several transitions ahead.

Uncertainty about future conditions also implies uncertainty about which actions to perform to solve problems or achieve goals. Performing the wrong action may not only fail to solve the problem, and thus squander valuable time and effort, but make the situation worse. It is to the benefit of the organism that it would solve the problem with a minimum of trial-and-error. High uncertainty implies poor control and thus low fitness. This in turn implies that during phylogenetic and ontogenetic development there will be a strong selective pressure for minimizing uncertainty, as organisms with higher uncertainty will lose the competition from those with a lower one.

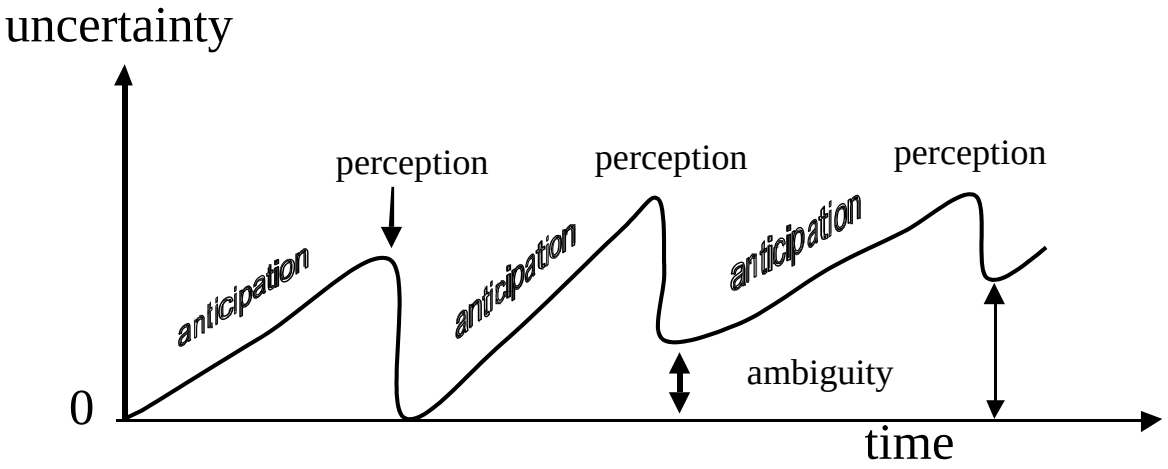
Phylogenetic evolution will lead to organisms with a more powerful in-built cognitive system. In higher organisms, this will take the shape of a brain, i.e. a crossroads where neural signals from sensors, sensing conditions, meet and combine, eventually moving out to effectors, performing actions. Some of the connections in this brain will be "hard-wired" by the genes, representing inherited instincts and reflexes. But since the environment is far too complex and changeful to be captured by rigid inherited rules, higher organisms will moreover develop the capability to learn, i.e. adapt their knowledge to their experience. Still, however good the knowledge, it will never be sufficient to predict with certainty. Anticipations need to be complemented by observations; feedforward must be corrected by feedback.

Uncertainty dynamics

An observation brings information into the system, and this by definition reduces uncertainty. This process is elegantly formalized in quantum mechanics: the state of a quantum particle, such as an electron, is represented by a wave of probability, which spreads out over space as it propagates. However, observing the position of the electron produces a sudden change in this state, which is called *the collapse of the wave function*. The observation process eliminates the spread or uncertainty about the particle's position, telling the observer precisely where it is located. But after this discontinuous collapse, the normal continuous evolution will again produce a gradual increase in uncertainty, as the wave spreads out again and the probability distribution becomes more and more diffuse.

In real cognitive systems, there is no clear separation between continuous spreading and discontinuous "collapse": both anticipation, which generates uncertainty as it extrapolates further into the future, and observation or perception, which reduces uncertainty as it confronts expectation with reality, take place more or less continously and simultaneously. Moreover, perception rarely *eliminates* uncertainty: usually the stimuli entering the cognitive system are too limited, noisy or ambiguous to produce an unequivocal interpretation. The remaining ambiguity may be resolved by anticipation, which will already have produced a "preselection" of potential or likely outcomes. The

interpretation that best fits with the constraints of both anticipation and perception is the one that the cognitive system will settle on. The more reliable the knowledge producing the anticipation, the higher the chances that this interpretation is the correct one.



But even if the organism cannot settle on a single, correct interpretation, any reduction in its uncertainty will significantly increase its chances of successfully tackling the situation. Thus, even a very incomplete or unreliable form of anticipation can spectacularly increase its chances for long-term survival and reproduction. For example, consider a situation that can be interpreted in a 1000 different ways, corresponding to 1000 different actions. Assume now that anticipation, extrapolating from previous observations, allows for 100 different possible outcomes, none of which is particularly more probable than any other. In terms of accurate prediction, this seems like a very weak mechanism, leaving a lot of uncertainty. But compared to the situation without anticipation, the number of potential actions to be considered has decreased tenfold, together with the probability that a fatal error would occur while trying out one of these actions. If an additional observation would moreover produce another tenfold reduction, the organism will meet with success after trying out on average 5 actions.

For a concrete illustration, imagine the very ambiguous stimulus produced by the play of light and shadow between the moving leaves of a large bush. Your interpretation of what may have caused this is likely to be very different depending on whether the bush is situated in your back yard or in an Indian jungle reputed for its tiger population. Although the probability that this stimulus was caused by a tiger is very small in both cases, in the second case it will appear much higher because your cognitive system is primed to look for tigers, and to some degree expects or anticipates that it may encounter one. In

this case, the wisest strategy is indeed to mentally prepare for a possible tiger attack, because the risks of ignoring that possibility are simply too great. In the first case, on the other hand, common sense tells us that it is not worth taking any specific precaution against tigers when checking out the origin of the perception.

We may conclude that the very limited sensory input that an organism has leaves a great amount of uncertainty about the actual state of a complex and changeful world, and about the appropriate actions to take in that situation. Any mechanism of anticipation or expectation that a priori reduces uncertainty will therefore have a large positive impact on the organism's capability for control, and thus on its fitness. The essence of cognition is precisely to produce such a mechanism that can complement or fill in incomplete data. To achieve optimal control, anticipation and perception, or feedforward and feedback, must go hand in hand, the one constantly correcting and extending the other.

Connectionist implementation

Given the above abstract, functional characterization of what a cognitive system must do, we will now try to formulate more concrete mechanisms through which it can perform this function. This is relatively straightforward if we interpret a condition-condition rule $A \rightarrow B$ as a connection or link between nodes A and B, with strength $P(A \rightarrow B)$. This maps the whole of the organism's knowledge onto a coherent, weighted and directed network, which is in general recurrent (i.e. contains cycles). Nodes represent elementary concepts or categories, i.e. the distinctions that the organism makes between classes of similar phenomena. Links represent the association or covariation that connects them together. The perception or anticipation of a concept then corresponds to an *activation* of the corresponding node, with a degree of activation a(A) proportional to the probability P(A) that the perceived or anticipated phenomenon actually is an instance of that concept.

Such a network is functionally equivalent to a so-called "neural" or "connectionist" network, that has been extensively used to simulate basic cognitive processes [Rumelhart, McLeod,]. I do not want to enter here into the discussion in how far such networks are realistic models of the actual neurons and synapses in the brain, which have various more complex features lacking in the simplified nodes and links of a connectionist network. Moreover, a concept is very unlikely to be represented by a single neuron, given that many neurons can be removed from someone's brain without observably damaging that person's cognitive capabilities. It is more likely that a concept is distributed over an "assembly" of neurons [Hawkins, 2004] that react with varying intensities to features associated with the concept, using a method called "population voting" [McCrone, 1999].

Learning algorithms

Whatever the precise correspondence between concepts and associations in the cognitive system, on the one hand, and neurons and synapses in the brain on the other hand, there seems to be a consensus that connectionist representations resemble the workings of the brain more than any of the other representations traditionally used in cognitive science/AI. This is seen most clearly in the basic neural learning mechanism postulated by Hebb []: in both connectionist representations and actual synapses, links that are successfully used are reinforced, in the sense that they will let more activation pass the next time they are invoked. Unused links, on the other hand, gradually become weaker.

This is a very general mechanism that can be motivated from simple physical assumptions outside the cognitive domain: any movement or flow, such as a river or a collective of animals walking across a landscape, will normally "erode" a distinct path, wearing off any obstructions to the flow. On the other hand, an unused path will normally be subjected to the accumulation of debris or the dissipation of structure so as to lose its differentiation from the environment. Thus, we can find analogs of the Hebbian adaptation mechanism in the carving out of valleys and canyons by rivers, the creation of trails by ants, and the breaking in of a car.

Neural network models have developed various refinements of this broad learning principle, such as "back propagation", to determine more accurately how much each contributing link should be strengthened or weakened depending on the overall succes of the process. Since backpropagation and related rules make specific assumptions about the architecture (feedforward, not recurrent) of the network and about the reinforcement function needed to measure the success of the outcome, we will limit ourselves to the simplest and most general refinement of the Hebbian principle: the Delta rule [McClelland & Rumelhart, 1988; McLeod, Plunkett & Rolls, 1998].

The delta rule compares the "external" or realized activation of a node B (e.g. received from perception or another independent mechanism) with its "internal" or "anticipated" activation (received from node A via the link $A \rightarrow B$). It then adjusts the link strength $w(A \rightarrow B)$ so as reduce the difference between the two. For example, if the anticipated activation value is lower than the actually observed one (underestimation), the link strength is incremented with a small amount, proportional to the difference. If it is higher (overestimation), the link strength is reduced correspondingly. This can be expressed by the following formula for the change in link strength, where $0 < \varepsilon \leq 1$ is a learning constant that determines how quickly the network adjusts to new experiences (and thus gradually forgets older experiences):

$$\Delta w(\mathrm{A} \rightarrow \mathrm{B}) = \varepsilon \,.(a_{ext}\,(B) - a_{int}\,(B))\,.a(A)$$

It can be shown that with this formula, as the number of experiences of perceiving A followed or not by B becomes large enough, the link strength will converge to a number proportional to the conditional probability of seeing B, given that A has occurred. The delta rule is mathematically equivalent to the Rescorla-Wagner rule which accurately predicts learning during classical conditioning experiments in animals and people. This makes it into an excellent candidate for a neurally plausible learning mechanism. Note also that the delta rule implements a negative feedback control loop, as it functions by compensating for any deviation from the desired outcome (accurate prediction).

A further advantage of both Hebbian and delta rules is that they function purely locally: they need information only about the activations at the start and at the end of the link to be adapted, not about the overall network architecture or specific problem to be solved. This allows the network to learn accurate anticipations in a very simple and natural way. Its only assumption about the structure of the world to be cognitively represented is a minimum of continuity, namely that phenomena that co-occurred regularly in the past are likely to continue doing so in the future. Even this assumption of continuity can be modulated by changing the learning constant, since a higher learning constant implies that earlier experiences contribute less to present expectations.

Spreading activation

Assuming that the cognitive system has learned an accurate connectionist representation of the co-occurrences or co-variations of the phenomena that it perceives, we now need to explain how it can use this network of links to make more complex predictions that involve several phenomena.

At any moment, several concepts will be activated to a greater or lesser extent, depending on present perception and anticipation from earlier perceptions. The distribution of activation over the nodes of the network defines the present *state* of the cognitive system. Since we assume that activation can vary continuously, there is a continuously infinite number of states. Even if activation could only take on two values, present or absent, the number of states would be 2^N, with N the number of nodes. Given a network with thousands, millions or even billions of nodes, this number is absolutely astronomical (2^{1000} is about 10^{301}, i.e. a 1 followed by 301 zeros). This illustrates the unimaginably powerful capacity of the cognitive system for representing different states of affairs, and the complexity of processing the corresponding cognitive states.

If just one concept A could be activated at a time, anticipation would be simple. We would just need to apply the matrix of conditional probabilities to

find which concept B is most likely to follow. When several concepts {A_1, A_2, ...} are active simultaneously with activations $P(A_1)$, $P(A_2)$, etc. , each will point to a number of other concepts {B_{11}, B_{12}, B_{13}, ...} with different conditional probabilities. Some of these concepts pointed at by different inputs will be the same, e.g. $A_1 \rightarrow B$, $A_2 \rightarrow B$. The question is then how to determine the probability for this shared anticipation B. The most common solution in connectionism is to *add* the activations propagated along each of the incoming links:

$$P(B) = \sum_i P(A_i).\ P(A_i \rightarrow B)$$

This is a simple and elegant model, that views activation like a kind of energy or fluid that spreads from node to node along the links, proportionately to the link strengths, but such that the total amount of activation in the system remains the same. It seems plausible that activation in the brain, which is carried by action potentials or concentrations of neurotransmitters, would propagate in such a way.

However, if we interpret activation as probability, this simple model runs into problems. Adding probabilities only makes sense if the alternatives are mutually exclusive: $P(A_1 \& A_2) = 0$. But in a self-organizing cognitive system, we cannot make any such assumption, since the nodes of the network do not represent separate states, but possibly overlapping categories that have in general been learned independently. For example, knowing that a phenomenon is a **pet** *and* a **bird**, we have a pretty good expectation that it may be a **parrot**, say P(**pet** & **bird** → **parrot**) = 0.30. However, the probabilities of it being a **parrot** knowing that it is either a **pet** or a **bird** are quite small, say: P(**pet** → **parrot**) = 0.05; P(**bird** → **parrot**) = 0.01. Using the additive rule in this case would lead us to determine P(**pet** & **bird** → **parrot**) = 0.05 + 0.01 = 0.06 « 0.30.

But this "super-additive" result is certainly not the general rule. Consider now the same probabilities for independent anticipations: P(**pet** → **fast_running**) = 0.05; P(**bird** → **fast_running**) = 0.01. However, P(**pet** & **bird** → **fast running**) = 0, as none of the fast running birds (such as ostriches) are pets. Yet, the first case in which the conjunction of two conditions leads to a more than additive probability estimation, seems like a more likely situation, given that confirmation by independent sources of evidence is generally taken as a sign that the hypothesis is quite strong.

In conclusion, the additive rule for combining conditional probabilities is a very coarse heuristic, that may be improved by some possible "super-additive" combination rule, e.g. of the form $P(B) = (P(A_1 \rightarrow B)^{1/k} + P(A_2 \rightarrow B)^{1/k})^k$, with $k > 1$. Various connectionist models use various non-linear functions to derive the actual activation of a node from the activation that enters via its incoming links. These functions typically use a threshold, so that the node only gets

activated if a sufficient amount of activation reaches it. We will not go into further details about these possible schemes. At present additive combination seems like a reasonable first approximation for what actually happens in the brain, and we will therefore use it as the paradigmatic mechanism.

Spreading activation, whether additive or not, will lead to more nodes being activated than those that got the initial activation.

+ cybernetic interpretation
    + anticipation = expectation = preparedness for action = feedforward
        + "priming" = preparing for action
            - when primed with an associated phenomenon (e.g. lion), recognition/decision-makin
about new phenomenon (e.g. tiger) is faster
            + partial pre-activation creates a bias for incoming activation
                - reaches threshold more easily
                - -> quicker and more accurate selection of the right interpretations/responses
                - less options need to be explored
        - makes it easier for certain types of action to be performed
        - feedforward is necessary in situations changing too rapidly or too complex for feedbac
control
    + no need for deterministic prediction, mere increased probability is already useful
        - conditional probability higher or lower than absolute probability
        - reduced conditional entropy -> lower entropy/uncertainty
        - reduces complexity of decision-making
        - reduces time to select adequate reaction

+ Expectation - cognitive control
  + memory prediction framework/ the anticipation control theory of mind
    + authors
        - Hebb, Neisser, McCrone, Hawkins
    + brain learns correlations/covariations through Hebbian rule
        - spreading activation/particle flow follows connections
        - when B is activated soon after A, strengthen connection A -> B
        + all non-used connections weaken exponentially
            - -> (non-)activation of A together with non-activation of B weakens A -> B
    + covarying elements determine invariant pattern
        + "transitively closed" subnetwork
            - sufficiently strongly interlinked so that activation propagates to all parts of the pattern
            - but dissipates outside of the pattern
            - cf. non-linear, "bootstrapping" clustering algorithm
        + activation of part of pattern tends to activate the whole
            - -> anticipation of as yet not perceived aspects from partial perception
        + subnetwork has single "label" or "name"
            - link to higher level node

+ the higher level network similarly forms closed subnetworks
  - -> recursive extraction of higher order covariance patterns
+ pattern can be spatial (simultaneous) or temporal (sequential)
  + spatial patterns = symmetric connections
    - transitive closure -> equivalence class -> simultaneous activation of all nodes in patterr
    + e.g face = spatial pattern
      - conjunctive combination of eyes, nose, mouth, hair, etc.
      + when some are perceived, the others are inferred
        - in either direction
  + temporal patterns = asymmetic connections
    - transitive closure -> partial order -> (branching) sequences of activation/expectation - further perception collapses branches to one sequence
    + e.g. melody = temporal
      - sequence of intervals between notes
      - when part of the melody is perceived, the remaining notes are anticipated one-by one
+ examples of psychological phenomena
  + pattern completion
    - actual input (e.g. eye & nose) is only small part of activated region (e.g. face)
    - spreading activation and closure determines the shape of the whole cluster
    + Gestalt perception
      - from correlated fragments (e.g. dots), a regular, closed figure is inferred (e.g triangle)
      - activation spreads to contiguous regions -> filling in of gaps
      + regular shapes (e.g. circle) are more common than specific irregular ones
        - -> anticipated, “filled in” fragments tend to follow regular outline
    + point light stimuli
      - person/object is recognized from covariation between movements of light dots
      + nothing is recognized if dots don’t move
        - insufficient information to infer implicit regularities
    + resolving linguistic ambiguity
      - context of preceding words/physical setting primes brain to select correct interpretatic of new word
    + false memory
      + brain remembers inferred fragments as well as perceived ones
        - e.g. list of words associated to “cold” -> subject remembers “cold” even though was not part of the list
        + poor witness reports
          - e.g. people remember black man & theft -> infer black man is thief -> rememb they saw a black man commit a theft
  + Need for variation to recognize invariants
    + static images become invisible
      - when projected so that they always stimulate the same cells in the retina
      - perceptual system needs variation to recognize anything

- otherwise, self-generated information (hallucination) is assumed

+ change blindness

- when actual input is changed indirectly without attracting attention to it (e.g. via flicke or very slowly), this is not noticed because the activated region remains continuous

+ Libet's half second

+ consciousness seems continuous though brain reaction has a delay of half a second t event/intention

- brain is constantly anticipating

+ recognizing faces

- implicit learning creates "average face template" as default expectation
- deviation from expectation -> uncontrolled -> a priori not as beautiful

+ getting to know the person

- deviation becomes predictable at a higher order
- -> "ugliness" disappears

+ specific face is stored at the higher level

- which only stores deviations from default

+ caricature = systematic exaggeration of deviations

+ easier to recognize than more realistic portrait

- because less confusion is possible with similar faces

+ also applies to other shapes

+ e.g. children's drawing of pig = more pink, more round, bigger snout, more curly ta ... than real pig,

- thus deviates more from "default animal" shape

+ esthetics

+ perception is beautiful when there is "flow"

- i.e. skills match challenges & challenges are high

- challenge = anticipating rest of partially perceived pattern
- skill = experience with recognizing similar patterns at different levels

+ higher skill (e.g. more experience, higher IQ)

- -> more challenge needed (more complex pattern)
- explains why "high brow", avant-garde art is appreciated only by minority of intellectua with broad art experience

+ parallel perception

- painting, photos, sculpture...
- scanned sequentially by eye movements, walking around, etc.
- recognizable figures -> allows anticipation of further scanning

+ personal "twist" -> anticipation meets with surprises

+ e.g. Magritte

- very recognizable archetypical symbols (clouds, trees, bodies, hats,)
- in unexpected combinations

+ sequential perception

- music, literature, film

+ rhythm, melody, plot ... are to large degree predictable

- but have sufficient surprising twists
+ on different levels
- e.g. sentences, events, episodes, character development
- e.g. rhythm, orchestration, theme, chorus, movement, symphony

+ consciousness
+ when perception deviates from anticipation
- = inconsistent with anticipated pattern
- then pass on to higher order control system
+ when highest order system is reached, subject becomes conscious of deviation
- focuses attention on deviation
+ orientation response = arousal, performing of actions that can bring in more information
- e.g. visual scanning, manipulation, moving around
- hippocampus is sollicited to produce a one-time, episodic memory of the event
+ event remains in memory for a long time
- the stronger the deviation, the longer

+ as long as anticipation is not falsified, processing happens automatically, implicitly unconsciously
- little is remembered
- e.g. driving a car for an experienced driver
+ not falsified ≠ correctly predicted
+ the signal may be too random to produce much prediction
- -> no significant anticipatory activation or priming
- in that case, no inconsistency or deviation is perceived
- but lower-order learning processes continue as usual
+ e.g. implicit learning of seemingly random sequences
- Cleeremans

+ control hierarchy
+ hierarchy
- covariation/correlation between varying input and anticipation-> invariant pattern (name, label, or concept)
- e.g. reference signal = difference between input and expectation
- when difference itself varies -> determines new varying perceptual input to higher level
- higher level produces anticipation
- if anticipation = perception -> new higher-order invariant pattern
+ cognitive control loop
- perception = perception
- goal = prediction as accurate as possible
- action = anticipate additional perception
- perturbation = unexpected events in the outside world
- information processing = compare anticipated with perceived phenomen
- error = difference between perception and anticipation
- action =

References

Chalmers DJ Facing up to the problem of consciousness, Journal of Consciousness Studies, 1995

Clark A. and Chalmers D. (1998):, The Extended Mind, *Analysis* 58, p. 7-19.

Clark, A. (1997). *Being There: putting brain, body, and world together again*, Cambridge, Mass., MIT Press.

George D. & Hawkins J. (2005): A Hierarchical Bayesian Model of Invariant Pattern Recognition in the Visual Cortex, Proceedings of the International Joint Conference on Neural Networks. (IJCNN 05)

Hawkins J. (2005): On Intelligence (Times books)

Lakoff, George; and Johnson, Mark. 1999. Philosophy in the flesh: the embodied mind and its challenge to western thought. New York: Basic Books.

McCrone J. (2000): Going Inside. A tour around a single moment of consciousness (Faber and Faber, London)

McLeod, P., Plunkett, K. & Rolls, E. T. (1998). *Introduction to connectionist modeling of cognitive processes*. Oxford, UK: Oxford University Press.

Neisser U. (1976): Cognition and Reality, San Francisco: Freeman.

O'Regan JK , A Noe (2001) A sensorimotor account of vision and visual consciousness Behavioral and Brain Sciences 24, 939–1031

Port, RF & T Van Gelder (eds.) (1996): Mind as motion: explorations in the dynamics of cognition (MIT Press, Cambridge, MA, 1996)

Rumelhart D.E. & J.L. McClelland (editors) (1986): *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1. MIT Press.

Searle JR 1980 Minds, brains, and programs, Behavioral and Brain Sciences,

Steels L. & Brooks R. (eds.) (1995): *The Artificial Life Route to Artificial Intelligence: Building Embodied Situated Agents* (Erlbaum).

Thelen E., LB Smith (1994): A dynamic systems approach to the development of cognition and action , MIT Press,

Varela, F.; Thompson, E. & Rosch, E. (1991). The Embodied Mind. Cambridge, MA: MIT Press.

Hawkins, J., Blakeslee, S., On Intelligence: How a New Understanding of the Brain will lead to Truly Intelligent Machines, Henry Holt and Company, 2004.

Alexandre, F., "Connectionist Cognitive Processing for Invariant Pattern Recognition", Proceedings International Conference on Pattern Recognition, 1996.

Körding, K.P., König, P., "Neurons with Two Sites of Synaptic Integration Learn Invariant Representations", Neural Computation, vol. 13, pp. 2823-2849, 2001.

Graziano, M.S., Gross, C.G., "Spatial maps for the control of movement", Current Opinions in Neurobiology, vol. 8, pp. 195-201, 1998.

Felleman, D., Van Essen, D.C., "Distributed hierarchical processing in the primate cerebral cortex", Cerebral Cortex, vol. 1, pp. 1-47, 1991.

Usher, M., Stemmler, M., Niebur, E., "The Role of Lateral Connections in Visual Cortex: Dynamics and Information Processing", In Lateral Interactions in the Cortex : Structure and Function , UTCS Neural Networks Research Group, Austin, TX, Electronic book, ISBN 0-9647060-0-8, 1996.

Sirosh, J., Miikkulainen, R., Bednar, J.A., "Self-Organization of Orientation Maps, Lateral Connections, and Dynamic Receptive Fields in the Primary Visual Cortex", Lateral Interactions in the Cortex : Structure and Function . UTCS Neural Networks Research Group, Austin, TX, Electronic book, ISBN 0-9647060-0- 8, 1996.

Stemmler, M., Usher, M., Niebur, E., “Lateral interactions in primary visual cortex: a model bridging physiology and psychophysics”, Science, vol. 269:5232, pp.1877-1880, 1995.