Poehlman, T.A., Uhlmann, E., Greenwald, A. G., and Banaji, M. R. (2006). 'Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity'. Unpublished MS.

Wilson, T. D., Lindsey, S., and Schooler, T. Y. (2000). 'A model of dual attitudes'. *Psychological Review,* 107.

attractor networks. Artificial neural networks (ANNs), sometimes referred to as **connectionist networks*, are computational models based loosely on the neural architecture of the brain. Over the past 20 years, ANNs have proven to be a fruitful framework for modelling many aspects of cognition, including perception, attention, learning and memory, language, and executive control. A particular type of ANN, called an *attractor network*, is central to computational theories of consciousness, because attractor networks can be analysed in terms of properties—such as temporal stability, and strength, quality, and discreteness of representation—that have been ascribed to conscious states. Some theories have gone so far as to posit that attractor nets are the computational substrate from which conscious states arise.

1. Artificial neural networks
2. Attractor dynamics
3. Relationship between attractor states and conscious states
4. Attractor networks and theories of consciousness

1. Artificial neural networks

ANNs consist of a large number of simple, highly interconnected neuron-like processing units. Each processing unit conveys an *activation level*, a scalar that is usually thought to correspond to the rate of neural spiking. Typically, activation levels are scaled to range between 0 (no spiking) to 1 (maximal spiking), and might represent the absence or presence of a visual feature, or the strength of belief in some hypothesis. For example, if the processing units—*units* for short—are part of a model of visual information processing, activity of a particular unit might denote the presence of the colour red at some location in the visual field. If the units are part of a model of memory, activity of individual units might instead denote semantic features of an item to be recalled.

Each unit receives input activation from a large number of other units, and produces output—its activation level—that is a function of its inputs. A typical function yields an output that grows monotonically with the weighted sum of the inputs. The weights, or strength of connectivity, can be thought of as reflecting the relationship among features or hypotheses. If the presence of feature A implies the presence of feature B, then the weight from A to B should be positive, and activation of A will tend to result in activation of B; if A implies the absence of B, the weight from A to B should be negative.

Units in an ANN can be interconnected to form two basic architectures: feedforward and recurrent. In a *feedforward architecture* (Fig. A25a), activity flows in one direction, from input to output, as indicated by the arrows. A feedforward network performs associative mappings, and might be used, for instance, to map visual representations to semantic representations. In a *recurrent architecture* (Fig. A25b), units are connected such that activity flows bidirectionally, allowing the output activity of a unit at one point in time to influence its activity at a subsequent point in time. Recurrent networks are often used to implement content-addres-

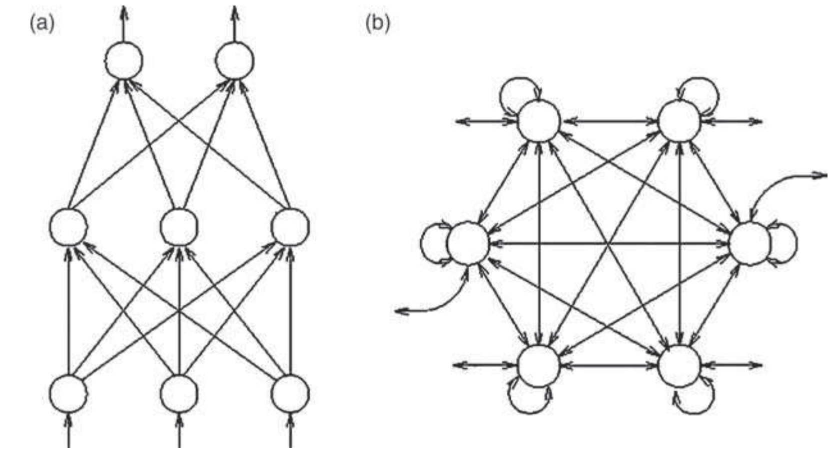

Fig. A25. (a) Feedforward architecture in which activity flows from the bottom layer of units to the top layer. (b) Recurrent architecture in which activity flows in cycles.

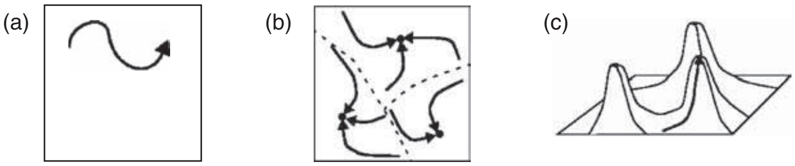

Fig. A26. (a) The activation state of a two-unit recurrent network can be depicted as a point in a two-dimensional space. If activation is bounded, e.g. to lie between 0 and 1, then the state lies within a box. The solid curve depicts the time-varying trajectory of the state, where the arrow represents the forward flow of time. (b) A state space with three attractors carved into attractor basins (dotted lines). (c) A harmony landscape over the state space with three attractors.

sible memories. The network is first trained on a set of items, and then when it is presented with a partial featural description of one item, network dynamics fill in the missing features. Both feedforward and recurrent networks can perform cued retrieval, but recurrent networks are more flexible in that they allow any subset of features to serve as a cue for the remaining features.

The activation state of a network with *n* units can be characterized as a point in an *n*-dimensional space, and temporal dynamics of a recurrent network can be described as a time-varying trajectory through this state space (Fig. A26a). An attractor network is a recurrent ANN whose dynamics cause the network state to converge to a fixed point. That is, given an input—which might represent a stimulus to be processed, or the output of another ANN—the dynamics of the network will cause the state to evolve over time to a stable value, away from which the state will not wander

2. Attractor dynamics

The states to which the net might evolve are called *attractors*. The attractors are typically sparse in the state space. (Technically, attractors can also be limit cycles—non-static, periodic trajectories—but attractor net dynamics ordinarily produce only point attractors.)

The state space of an attractor net can be carved up into *attractor basins,* regions of the state space in which all starting points converge to the same attractor. Figure A26b depicts a state space with three attractor basins whose boundaries are marked by dotted lines, and some trajectories that might be attained within each attractor basin.

Attractor dynamics are achieved by many neural network architectures, including Hopfield networks, harmony networks, Boltzmann machines, adaptive resonance networks, and recurrent back-propagation networks. To ensure attractor dynamics, these popular architectures require symmetry of connectivity: the connection weight from processing unit A to unit B must be the same as the weight from B to A. Given this restriction, the dynamics of the networks can be characterized as performing local optimization—minimizing energy, or equivalently, maximizing harmony. Consider the attractor state space of Fig. A26b, and add an additional dimension representing harmony, a measure of the goodness of a state, as shown in Fig. A26c.

The attractors are at points of maximum harmony, and the network dynamics ensure that harmony is non-decreasing. Because the net is climbing uphill in harmony, it is guaranteed to converge to a local optimum of harmony. The input to an attractor net can either specify the initial state of the net, or it can provide biases—fixed input—to each unit; in the latter case, the biases reshape the landscape such that the best-matching attractor has maximum harmony, and is likely to be found for a wide range of initial network states.

The connection strengths (including biases) in the network determine the harmony landscape, which in turn determines the attractors and the shape of the attractor basins. When a set of attractor patterns are stored in a net, gang effects are typically observed: the shapes of attractor basins are influenced by the proximity of attractors to one another (Zemel and Mozer 2001).

In traditional attractor nets, the knowledge about each attractor is distributed over the connectivity pattern of the entire network. As a result, sculpting the attractor landscape is tricky, and often leads to spurious (undesired) attractors and ill-conditioned (e.g. very narrow) attractor basins. To overcome these limitations, a localist attractor net has been formulated (Zemel and Mozer 2001) that consists of a set of state units and a set of attractor units, one per attractor. Each attractor unit draws the state toward its attractor, with the attractors closer to the state having a greater influence. The localist attractor net is easily configured to obtain a desired set of attractors. The dynamics of a localist attractor net, like its distributed counterpart, can be interpreted as climbing uphill in harmony.

Attractor nets can also be conceptualized from a probabilistic perspective. If the net has intrinsically stochastic dynamics (e.g. Boltzmann machines, or back-propagation networks with added noise), each point in the state space can be characterized in terms of the probability of reaching each attractor from that point—a discrete probability distribution over attractors. The points far from any attractor have a nearly uniform distribution (maximum entropy), whereas the attractors themselves are represented by a distribution with probability 1.0 for the attractor and probability 0.0 for any other attractor (minimum entropy). This conceptualization allows for one to abstract away from neural net representations and dynamics, and to characterize the dynamics as entropy minimization (Colagrosso and Mozer 2004).

3. Relationship between attractor states and conscious states

Theorists have identified certain properties that are claimed to be prerequisites or characteristics of conscious mental states. Attractors share these properties, as we elaborate here.

Conscious perceptual states have been conceived of as interpretations of noisy or ambiguous sensory input (Marcel 1983). For example, the Necker cube admits two possible interpretations, and perceptual awareness flips between these interpretations (see *multistable perception). Searle (1992) focuses on interpretation in terms of pre-existing categories. One can conceive of attractors as interpretations or learned categories, and the dynamics of an attractor net as mapping a noisy or partial input to the most appropriate interpretation. Attractor dynamics are highly non-linear: two similar initial states may lead to distant attractors. This type of non-linearity allows two similar inputs to yield distinct interpretations. The Necker cube is an extreme case in which a single input—lying on the boundary between two attractor basins—can lead to two different interpretations. (Many attractor nets assume intrinsic noise to break symmetry for ambiguous inputs.)

Conscious states have been characterized as high-quality representations (Farah 1994, Munakata 2001). The notion of quality is ill defined, but essentially, a high-quality representation should be capable of triggering the correct representations and responses further along the processing stream; in the terminology of the consciousness literature, such a representation is *accessible*. Quality is not an intrinsic property of a representation, but comes about by virtue of how that representation affects subsequent processing stages, which in turn is dependent on whether past learning has associated the representation with the appropriate effects. From this definition, attractors are high quality. Attractors come into existence because they correspond to states the system has learned about in the past. An attractor net cleans up a noisy input, yielding a pattern that corresponds to a previously experienced state. Because of this past experience, later stages of processing receiving input from the attractor net are likely to have learned how to produce appropriate responses to attractor states. When cognitive operations involve multiple steps, the quality of a representation is critical: without the sort of clean-up operation performed by an attractor net, representations degrade further at each step (Mathis and Mozer 1995).

Temporal stability of neural states is often associated with consciousness (e.g. Taylor 1998). Attractors have the property of temporal stability. Once the dynamics of the attractor net lead to an attractor, the state the state of the network persists until the network is reset or is perturbed by a different input.

Conscious states are generally considered to be *explicit* (e.g. Baars 1989, Dehaene and Naccache 2001), meaning that they are instantiated as patterns of neural activity, in contrast to *implicit* representations, which are patterns of connectivity. An attractor net encodes its attractors implicitly, but the current attractor state is explicit.

Conscious states might arise at the interface between sub-symbolic and symbolic processing (Smolensky 1988, Cleeremans and Jiménez 2002). From a connectionist perspective, perceptual processes are intrinsically sub-symbolic, but yield representations of object identities and categories that subserve subsequent symbolic processing. This view fits in well with the fact that attractor nets typically map a continuous activation space to a discrete set of alternatives (Fig. A26b), which can be viewed as a mapping from sub-symbolic to symbolic representations. If conscious states are indeed symbolic, then they should be all-or-none. Studies have indeed suggested discrete, all-or-none states of consciousness (Sergent and Dehaene 2004), although others consider consciousness to be a graded phenomenon (Farah 1994, Munakata 2001, Cleeremans and Jiménez 2002).

4. Attractor networks and theories of consciousness

Grossberg's *adaptive resonance* theory, proposed in 1976, describes an attractor network that achieves resonant states between bottom-up information from the world and top-down expectations. Grossberg (1999) subsequently made the claim that conscious states are a subset of resonant (attractor) states. From *functional brain imaging data, evidence is also consistent with the notion that conscious states arise from resonant circuits linking temporal, parietal, and prefrontal cortical areas (Lumer and Rees 1999).

Many computational theories of consciousness have argued that attractors have the right functional characteristics to serve as the computational *correlate of consciousness. Rumelhart et al. (1986) and Smolensky (1988) first proposed that conscious mental states may correspond to stable states of an attractor network. Like subsequent theorists, they envision an attractor net defined over multiple cortical regions thereby able to capture global cortical coherence. Farah et al. (1993) describe attractor nets as allowing stimuli to be integrated into a global information-processing state which corresponds with consciousness. Other theorists are less explicit in describing attractor nets, yet focus on key properties of attractor nets, such as self-sustaining activation patterns (Dehaene and Naccache 2001), dynamic competitions among coalitions of neurons (Crick and Koch 2003), and non-linear bifurcations in neural activity (Sergent and Dehaene 2004). Although many theories simply postulate that stable, high-quality representations—such as attractors—are associated with awareness, some models to show that accessibility and reportability is an emergent property of such representations (Mathis and Mozer 1996, Colagrosso and Mozer 2004).

MICHAEL C. MOZER

Baars, B. (1989). *A Cognitive Theory of Consciousness*.

Colagrosso, M. D. and Mozer, M. C. (2005). 'Theories of access consciousness'. In Saul, L. K., Weiss, Y. and Bottou, L. (eds.) *Advances in Neural Information Processing Systems, 17*.

Cleeremans, A. and Jiménez, L. (2002). 'Implicit learning and consciousness: a graded, dynamic perspective'. In French, R. M. and Cleeremans, A. (eds) *Implicit Learning and Consciousness*.

Crick, F. and Koch, C. (2003). 'A framework for consciousness'. *Nature Neuroscience, 6*.

Dehaene S. and Naccache L. (2001). 'Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework'. *Cognition, 79*.

Farah, M. J. (1994). 'Visual perception and visual awareness after brain damage: A tutorial overview'. In Umiltà, C. and Moscovitch, M. (eds) *Attention and Performance XV: Conscious and Nonconscious Information Processing*.

——, O'Reilly, R. C., and Vecera, S. P. (1993). 'Dissociated overt and covert recognition as an emergent property of a lesioned neural network'. *Psychological Review, 100*.

Grossberg, S. (1999). 'The link between brains, learning, attention, and consciousness'. *Consciousness and Cognition, 8*.

Lumer, E. D. and Rees, G. (1999). 'Covariation in activity in visual and prefrontal cortex associated with subjective visual perception'. *Proceedings of the National Academy of Sciences of the USA, 96*.

Marcel, A. J. (1983). 'Conscious and unconscious perception: an approach to the relations between phenomenal experience and perceptual processes'. *Cognitive Psychology, 15*.

Mathis, D. A. and Mozer, M. C. (1995). 'On the computational utility of consciousness'. In Tesauro, G. et al. (eds) *Advances in Neural Information Processing Systems, 7*.

—— —— (1996). 'Conscious and unconscious perception: a computational theory'. In Cottrell, G. (ed.) *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society*.

Munakata, Y. (2001). 'Graded representations in behavioral dissociations'. *Trends in Cognitive Sciences, 5*.

Rumelhart, D. E., Smolensky, P., McClelland, J. L., and Hinton, G. E (1986). 'Schemata and sequential thought processes in PDP models'. In Rumelhart, D. E. et al. (eds) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 2.

Searle, J. R. (1992). *The Rediscovery of the Mind*.

Sergent, C. and Dehaene, S. (2004). 'Is consciousness a gradual phenomenon? Evidence for an allornone bifurcation during the attentional blink'. *Psychological Science, 15*.

Smolensky, P. (1988). 'On the proper treatment of connectionism'. *Behavioral and Brain Sciences, 11*.

Taylor, J. G. (1998). 'Cortical activity and the explanatory gap'. *Consciousness and Cognition, 7*.

Zemel, R. S. and Mozer, M. C. (2001). 'Localist attractor networks'. *Neural Computation, 13*.

autism. The term 'autism', from the Greek word meaning self, was first applied to children with notably abnormal social development by Leo Kanner (1943) and Hans Asperger (1944)—who took the term from Bleuler's (1911) description of withdrawal and self-absorption in *schizophrenia.

Autism is a neurodevelopmental disorder diagnosed by the presence of qualitative social and communicative impairments and restricted and repetitive interests and activities, manifest before age three. A child with autism may be silent and socially aloof, and line up toys repetitively; an adult with high-functioning autism may be verbally fluent in a pedantic monologuing way, socially over-eager in a gauche way, and repetitive in his narrow interest in, say, electricity pylons. Since the range of manifestations of this 'triad' of impairments varies with age and ability, the notion of *autism spectrum disorders* (ASD) has become popular. This spectrum includes *Asperger's syndrome*, in which the key features of autism are present but there is no general intellectual or language delay. ASDs are not as rare as once thought; perhaps as many as 1 in 1000 individuals may have autism, and almost 1 in 100 may have an ASD. The majority of affected individuals are male, and many people with autism also have general intellectual impairment. Search for the causes of ASD continues; while there is a strong genetic component (autism is among the most highly heritable psychiatric disorders), no specific genes or brain basis have as yet been identified.

Prominent among psychological accounts of the distinctive behavioural impairments in ASD is the idea of deficits in *theory of mind* or *mentalizing*: the everyday ability to attribute beliefs, desires, and other mental states to self and others in order to explain and predict behaviour. There is now overwhelming evidence from a variety of simple tests that most people with ASD have difficulty knowing what others think and feel (reviewed