HISTORY AND PHILOSOPHY OF NEURAL NETWORKS

J. Mark Bishop
Department of Computing, Goldsmiths, University of London, New Cross, London

Keywords: Artificial Intelligence, Connectionism, Embodiment, Neural Networks, Mind

Contents

Summary

This chapter conceives the history of neural networks emerging from two millennia of attempts to rationalize and formalize the operation of mind. It begins with a brief review of early classical conceptions of the soul, seating the mind in the heart; then discusses the subsequent Cartesian split of mind and body, before moving to analyze in more depth the twentieth century hegemony identifying mind with brain; the identity that gave birth to the formal abstractions of brain and intelligence we know as ‘neural networks’.

The chapter concludes by analyzing this identity - of intelligence and mind with mere abstractions of neural behavior - by reviewing various philosophical critiques of formal connectionist explanations of ‘human understanding’, ‘mathematical insight’ and ‘consciousness’; critiques which, if correct, in an echo of Aristotelian insight, suggest that cognition may be more profitably understood not just as a result of [mere abstractions of] neural firings, but as a consequence of real, embodied neural behavior, emerging in a brain, seated in a body, embedded in a culture and rooted in our world; the so called 4Es approach to cognitive science: the *Embodied*, *Embedded*, *Enactive*, and *Ecological* conceptions of mind.

1. Introduction: The Body and the Brain

What force directs these actions? Where lies the body-magic that brings forth my world? Where sits this mind? If science can fully respond to these questions then it might be possible one day to simulate the human ability to act mindfully, with intelligence, on computer.

For much of the twentieth century the dominant paradigm of intelligence seated the mind in the brain; thus, if computers can model the brain then, theory goes, it ought to be possible to program computers to act intelligently. In the latter part of the twentieth century this insight - that intelligence is grounded in the brain - fuelled an explosion of interest in computational "neural networks" : high fidelity accurate simulations of the brain (cf. 'computational neuroscience') or engineering approximations used to control intelligent machines (connectionism). However, the view that intelligence is rooted solely in the brain is a relatively modern one and one that, in recent years, is being challenged by embodied approaches to artificial intelligence; a perspective that, in turn, can be traced back to the classical era.

The classical view of the mind was encompassed by the earlier notion of soul, leading Aristotle (in the *De Motu Animalium*, 350 B.C.E.) to famously enquire "how the soul moves the body, and what is the origin of movement in a living creature". In stark contrast to modern notions identifying mind with brain, Aristotle posited "the heart is the seat of the senses" - the sensorium - and the controller of both voluntary and involuntary movement; unlike, say, Alcmaeon and Hippocrates, in Aristotle's metaphysics of movement and tripartite division of the soul (into the appetitive, sensitive and rational parts; localized respectively in the liver, heart, and brain) there was simply no place for the brain in the casual chain that ultimately gave rise to animal behavior.

Such classical ideas had influence well into the renaissance period until, in 1649, Descartes laid the foundations of a new - dualist - division of body (*res extensa*) and the immaterial soul/mind (*res cogitans*); the mind interacting with the material brain [the organ which controlled the body] - at an interface famously located in the pineal gland. Around this time support for this view of the brain - as the organ that controlled behavior - found empirical support in the work of a contemporary of Descartes, the English physician Thomas Willis, who in 1667 began to identify links between the physical structure of the brain and certain pathological behaviors (e.g. epilepsy and other convulsive diseases).

Emerging from what later became known as the British Empiricist school of philosophy, John Locke was perhaps the first philosopher to define the self through a 'continuity of consciousness':that conscious thinking thing, (whatever substance, made up of whether spiritual, or material, simple, or compounded, it matters not) which is sensible, or conscious of pleasure and pain, capable of happiness or misery, and so is concerned for itself, as far as that consciousness extends.

In addition, in his 1690 'essay concerning human understanding' Locke famously suggested that at birth the mind was to be considered a blank slate ('*tabula rasa*'):

> Let us then suppose the mind to be, as we say, white paper, void of all character, without any ideas. How comes it to be furnished? I answer, in one word, from experience.

Thus, contrary to prevailing Cartesian philosophy (which held the basic logical propositions are innate), Locke maintained that we are born without innate ideas, and that all 'knowledge' comes from 'experience'; that knowledge is causally dependent on experience and that knowledge is justified solely by experience. Subsequently building on these intuitions in a supplementary chapter of the fourth edition of his essay, in 1690 Locke introduced the notion of an 'Association of Ideas' to label the principle accounting for the mental peculiarities of individuals. (In the philosophy of mind a 'theory of mind' typically attempts to explain the nature of ideas, concepts and other mental content in terms of the 'cognitive states' of underlying 'cognitive processes'. A cognitive state can thus encompass knowledge, understanding, beliefs etc. In a 'representational theory of mind' the cognitive states are conceived in terms of relations to 'mental representations' (which have content). In this view the underlying cognitive processes are simply understood in terms of 'mental operations' on the mental representations.). In Locke's 'associationist theory of mind' this *association of ideas* - or associationism as it later became known - suggested that the mind is organized, at least in part, by principles of association and that items that 'go together' in experience will 'go together' in thought; subsequently David Hume refined Locke's generic notion of 'going together by association' by reducing it to three core empirical principles: identity, contiguity in time and place, cause and/or effect.).

The associated ideas ('representations') could be memories, ideas, images, thoughts etc., with complex ideas being constructed from 'simples' and simple ideas being derived from sensations. Such raw sensations/perceptions were not governed and defined by principles of association, but were externally caused by things 'outside the head', which for Hobbes, Hume and Locke meant objects of the world. Hence for some time associationism was closely linked with the broad British Empiricist movement in philosophy.

Building on earlier ideas from Galileo and Descartes, Locke's epistemology famously identified objects [of reality] by their primary and secondary qualities. Examples of primary qualities include: solidity; extension; figure; number; motion; rest etc. Clearly the primary qualities have a direct link to their bearer; a primary quality says something about its bearer. E.g. If an object instantiates the primary quality of rest, then the object [of which it is a property] must be at rest. In this manner Locke suggested such primary qualities are essential to their bearers and are intrinsic qualities of their bearer and hence are fundamentally independent of the perceiving mind.

Conversely, Locke hypothesized the 'secondary qualities' to be the properties [of objects] that produce sensations in observers; things such as color, sound, taste, smell etc. In this conception secondary qualities are the powers of objects - by configurations of their primary qualities - to cause experience in subjects; the effect objects have on some people. For Locke primary qualities exist in the world but secondary qualities exist only in the mind of the perceiver; there is no blueness or sweetness in the world, only extension in motion.

Thus Locke's associationism entails (i) compounding processes, where complex items are formed from simple items; (ii) decompositional processes, where complex items are broken down into their simple elements and (iii) sequencing processes, where associations follow one another (e.g. in time). In sequencing processes groups of items (e.g. memories) follow one another in one of two ways: (a) by intrinsic association, whereby some items have a natural connection (i.e. a connection independent of the observer) with each other. E.g. chili pepper and 'hotness' and (b) by extrinsic association, whereby some items have an observer dependent connection (either voluntarily or by happen chance). E.g. If the first person one fell madly in love with had long red hair, then one might thereafter associate these qualities with beauty, sex and love.

In his essay Locke suggested that such extrinsic associations had three properties: (i) they are either voluntary or happen by chance; (ii) the strength of the 'impression' of ideas can reinforce the association (i.e. powerful ideas may be forever linked in the mind. cf. perception of beauty) and (iii) some items will 'go together' more easily than others; thus some will find mathematical associations easy, some artistic etc. Thus, although at birth the mind is a blank slate empty of ideas, individuals may find some 'ideas' easier to associate than others. These 'ideas' are conceived as mental representations; in this way we entertain particular ideas when in particular mental states. Thus the Scottish philosopher, historian and economist David Hume famously suggested that mental processes are merely sequences of such associated mental ideas. But what are such mental 'ideas' actually 'about'?

To what do 'ideas' actually refer? This, of course, is the problem of intentionality: 'how do mental ideas connect with [become to be about] things in the world?' Hume's response was to suggest that mental ideas are fundamentally representations 'like images'; thus positing a pictorial resemblance between idea and world. Unfortunately Hume's pictorial mechanism is not without problems. Firstly, it is both too general (not all ideas are like pictures; cf. justice) and not general enough (consider a picture of Eiffel Tower; this image may look like the Eiffel Tower, from one perspective, but that image is not necessary to the notion of the Eiffel Tower; the tower that Gustave Eiffel built); secondly, it provides no account of mental reference (resemblance is not sufficient for representation; a cartoon may look more like its creator than its subject, but can still represent the subject); thirdly, it offers no account of truth and falsity. Contra Tractatus Wittgenstein [2.17] "What the picture must have in common with reality in order to be able to represent it after its manner - rightly or wrongly - is its form of representation", images are not propositions; images [in themselves] are neither true or false.

Furthermore, although Hume's associationism - where mental/cognitive processes are simply defined as associations between representations - can easily accommodate [folk] psychological explanations of mind, it fundamentally lacks a workable account of representational content. i.e. what is it about a 'representation of a dog' that constitutes it as representing a dog rather than a banana (or anything else), rather than not being representational at all? Even in the present era [according to the American philosopher Mark Bickhard] the problem of "accounting for representational content is the central issue in contemporary naturalism: it is the major remaining task facing a naturalistic

conception of the world. Representational content is also the central barrier to contemporary cognitive science and artificial intelligence: it is neither possible to understand representation in animals nor to construct machines with genuine representation given current (lack of) understanding of what representation is" (Bickhard, 1993).

The preliminary theoretical base for a neural conception of mind (and hence also for contemporary neural networks) was independently proposed by Alexander Bain (1873) and William James (1891). Central to their work is the notion that both thoughts and body activity resulted from neuronal processes in the brain; for Bain every possible activity required the firing of a distinct set of neurons. At the time the scientific community was skeptical of Bain's ideas because they appeared to require an inordinate number of neural connections within the brain (albeit it is increasingly apparent that the brain is an exceedingly complex organ and that the same brain 'hardware' can process multiple problems and multiple inputs - see Section 10.4.3).

In some ways James's theory was similar to Bain's, however in James's model he suggested that memories and actions resulted from electrical currents flowing between neurons; this was perceived as a significant advantage as, by focusing on the flow of electrical currents between neurons, it did not require as many distinct neural groups to store memories or motivate action.

1.1. William James and Neural Associationism

In "Minds, Brains, Computers" Harnish centrally views William James "The Principles of Psychology" (James,1891) as suggesting that thinking (indeed all aspects of conscious life) had the following key properties:

- Thinking is conscious.
- Thinking is open to introspective examination.
- Thinking is private; 'my thought belongs with my other thoughts and your thought belongs with your other thoughts'.
- Thinking 'flows like a stream': a metaphor that gives rise to the idea of the 'stream of consciousness'; a concept with significant resonance in twentieth century literature, most famously in the works of Virginia Woolf and James Joyce.
- Thinking is 'about something' (i.e. it is fundamentally "intentional").
- Thinking is an evolved function (i.e. it is not a 'gift from god').

For James the key question in psychology is 'how does the mind solve the problem of what to think next?' The answer James arrived at was that thinking fundamentally operates using two general 'principles of association' which James sought to explain on a quasi-neurological basis:

- Principle 1:*When two elementary brain processes have been active together or in immediate succession, one of them, on reoccurring, tends to propagate its excitement into the other; a mechanism remarkably similar to the learning scheme outlined in 1949 by the Canadian psychologist Donald Hebb* (1949) (see Section 3.1)*)*
- Principle 2: *The amount of activity at any given point in the brain cortex is the sum tendencies of all the other points that discharge into it – a similar mechanism to*

that described in 1943 by McCulloch & Pitts (1943) *in their mathematical model of neural firing* (see Section 2.1) - *such tendencies being proportionate:*

- *to the number of times the excitement of each point may have accompanied the point in question;*
- *to the intensity of the excitements;*
- *to the absence of any rival point, functionally disconnected with the first, into which the discharges may have been diverted.*

Using these principles James offers the following explanation of three types of associative mental processes involved in spontaneous thought: unrestricted association; partial association and focused association.

Total association: in which there is unrestricted association between previous events and/or arbitrary concepts. e.g. James famously describes 'The memory of a walk followed by a romantic dinner'.

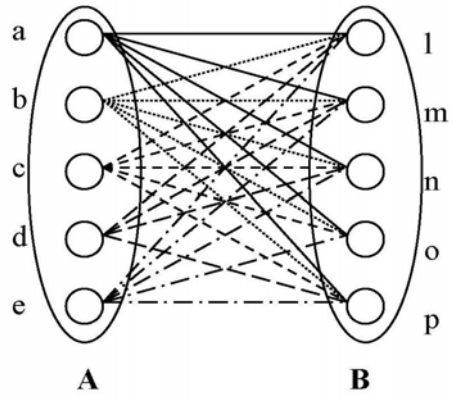

Figure 1. Total Association

Let **A** be the memory of the walk where the pattern of excitement distributed across neurons $[a, b, c, d, e]$. Let **B** be romantic thoughts where the pattern of excitement distributed across neurons $[l, m, n, o, p]$. To store this unrestricted total association James suggests the memories A and B must "vibrate in unison" hence '**A** must excite **B**' and '**B** must excite **A**', see Figure 1.

Partial association in which only some of the past experiences have associated consequences; but 'why are some memories linked and not others?' James suggests that, "in no revival of past experience are all the items of thought equally operative in determining what the next thought shall be. Always some ingredient is prepotent over the rest." For James the item which is prepotent is the one which is most in our interest, "some one brain-process is always prepotent above its concomitants in arousing elsewhere".

James claims partial recall is the most common form of association-based recall and occurs when only some of past experiences have the required 'associated consequences'. To determine what experiences are [partially] associated together James outlines four principles; these are:

- **Habit:** the more often something done the more likely it is to be associated.
- **Recency:** more recent events are more likely to be recalled.
- **Vividness:** the more intense an experience the more likely it is to be recalled.
- **Emotional congruity:** similar emotion backgrounds are more likely to be associated together; hence feeling miserable makes it more difficult to recall times of joy.

At any time *the strongest principle of association* is that which pertains.

To see how these principles of association work imagine if one is thinking **A** and **A** is associated with **B**, (by say habit), then one will subsequently think **B** *unless a stronger principle applies*.

Focalized recall / association by similar sequences: in which not all memories are obviously associated ('go together'). As an example James considers: 'How might thoughts of a gas flame lead to thoughts of football?' James unfolds the process in this way: think of the gas-flame which conjures up visions of the moon [via a shared 'similarity of color'; pale-whiteness]; this, in turn, is linked to the thoughts of football [via 'similarity of shape'; roundness]. So via focused recall thought can move from **A** to **B** (from gas-flame to football) even though neither gas-flame nor football have any properties in common by themselves.

Contrasting the above examples of 'spontaneous thought' James also offers explanation for voluntary thought; a subject often thought problematic for associationists. James addresses the issue by offering an associationist account of the process of 'recall of a forgotten thing' (and claims a similar mechanism underlies means-end analysis).

To recall a forgotten item James envisages a situation where two groups of three memories $[a,b,c]$ and $[l,m,n]$ are fully associated (& interconnected) with the 'forgotten item' **Z**.

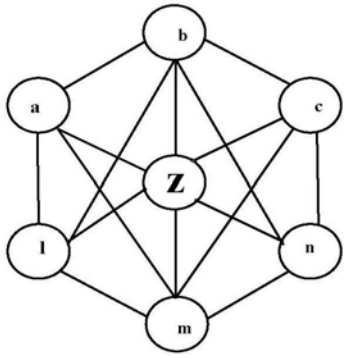

Figure 2. To recall a memory

By the process of total association activation of $[a,b,c]$ will eventually propagate to activate $[l,m,n]$ which together $[a,b,c, l,m,n]$ will serve to activate the forgotten memory **Z**, see Figure 2.

Means-end analysis is performed in a similar way. If $[a,b,c]$ is the goal states, its activation will eventually propagate to excite suggestion-states $[l,m,n]$ which together will excite the solution $\mathbf{Z}$, albeit in his analysis Harnish points out that James never considers if all reasoning is performed via means end or even if his version of means-end analysis will work in practice; for example, Harnish suggests that balancing a bank account would be difficult to describe in this way.).

1.2. The Neuron: Fine Grain Structure of the Brain

In "Aristotle's laptop: the discovery of our informational mind". Aleksander and Morton (2012) suggest research into the building blocks of the brain effectively began with the discovery that all living things are made of cells. Thus, 1838 the German botanist Matthias Schleiden first identified 'cell like' structures in plants; an idea later extended by physiologist Theodor Schwann who observed similar 'cellular structures' in the organs of animals; a discovery which, in turn, led to the development of 'cell theory': the idea that all living materials - whether in plants or animals - are composed of cells. In the context of the brain this posed the question: 'are the cells of the brain just a way of creating a fused functional structure or is there something special that the cells of the brain do (at the individual cell level) which is different to the function of other cells?'

Central to the search to resolve this issue was the ability to stain tissue such that it shows up better under microscopes (e.g. see Figure 3), research pioneered by the Bohemian anatomist Johannes Evangelista Purkinnje. Johannes worked at the University of Prague where he successfully made the then thinnest known slices of brain tissue; to make the slices Purkinnje was amongst the first to use a device called a microtome; a tool to cut extremely thin slices of material.

Figure 3. An example brain slice clearly showing neuron-bodies, the dendritic-tree and axon-fibers

Bibliography

[1] Abbot, L., (1990), Learning in Neural Network Memories, Network: *Computation in Neural Systems* 1, pp. 105-122. [Offers a critique of neural modeling in classical connectionist models of low level cognition].

[2] Auer, P, Burgsteiner, H. & Maass, W., (2002), Reducing communication for distributed learning in neural networks, in: J. R. Dorronsoro (Ed.), (2002), Proc. of the International Conference on Artificial Neural Networks ICANN 2002, Vol. 2415 of Lecture Notes in Computer Science, pp. 123-128, Springer.. [Auer suggests that to approximate a given input/output behavior F on particular functions of time (or spike trains), it is simply necessary to randomly pick some sufficiently complex reservoir of recurrent circuits of spiking neurons, and then merely adapt the weights of a single pool of feedforward spiking neurons to approximate the desired target output behavior].

[3] Ackley, D.H., Hinton, G,E & Sejnowski, T,J., (1985). A Learning Algorithm for Boltzmann Machines. Cognitive *Science*: 9 (1), pp. 147?169. [First description of Boltzmann machine learning algorithm for neural networks with hidden layers].

[4] Ahmad, S., (1991), VISIT: An efficient computational model of human visual attention, PhD Thesis, University of Illinois, USA. [Another critique of neural modeling in classical connectionist models of low level cognition.]

[5] Aleksander, I, (1970), Microcircuit learning networks: hamming distance behavior, *Electronics Letters* 6 (5). [Early introduction to the idea of using small microcircuit devices to implement Bledsoe and Browning's n-tuple recognition systems]

[6] Aleksander, I. & Stonham, T.J., (1979), "Guide to Pattern Recognition using Random Access Memories", *Computers & Digital Techniques* 2 (1), pp. 29-40, UK. [Early review of the n-tuple or weightless network paradigm.]

[7] Aleksander, I, Thomas, W.V. & Bowden, P.A., (1984),"WISARD a Radical Step Forward in Image Recognition", *Sensor Review*, UK. [A description of a real-time vision system built using weightless networks.]

[8] Aleksander, I & Morton, H., (1995), *An Introduction to Neural Computing*, Cengage Learning EMEA. [Introductory text on neural networks.]

[9] Aleksander, I & Morton, H., (2012), *Aristotle's laptop: the discovery of our informational mind*, World Scientific Publishing, Singapore. [Research text outlining Aleksander and Morton's idea of the 'informational mind'.]

[10] Bain, A, (1873), *Mind and Body. The theories of their relation*, D.Appleton & Co, NY. [In which Bain suggests that our thoughts and body activity result from interactions among neurons within the brain.]

[11] Barto, A.G., Sutton, R. S. & Anderson, C. W., (1983), Neuron like adaptive elements that can solve difficult learning control problems, *IEEE Transactions on Systems, Man and Cybernetics* SMC:13: pp. 834-846.[The classic introduction to reinforcement learning.]

[12] Barnden, J., Pollack, J. (eds.), (1990), High-Level Connectionist Models, Ablex, Norwood NJ, USA. [Edited volume which includes critiques of neural modeling in classical connectionist models of high level cognition.]

[13] Beer, R. D., (1995), On the dynamics of small continuous-time recurrent neural networks, *Adaptive Behaviour* 3(4), pp. 469-509. [A key introduction to Continuous Time Recurrent Neural Networks.]

[14] Benacerraf, P., (1967), God, the Devil & Gödel, *Monist* 51.[Benacerraf's response to Lucas's Gödelian argument against materialism.]

[15] Bickhard, M. H. (1993). Representational Content in Humans and Machines. *Journal of Experimental and Theoretical Artificial Intelligence* 5, pp. 285-333.[Mark Bickard highlighting the problem of symbol grounding in cognitive science.]

[16] Bickhard, M. H. & Terveen, L.,(1995). *Foundational Issues in Artificial Intelligence and Cognitive Science*, North Holland. [Introduction to Mark Bickhard's interactionist account of cognition.]

[17] Bishop, J.M. (1989), Stochastic Searching Networks, Proc. 1st IEE Int. Conf. on Artificial Neural Networks, pp. 329-331, London, UK. [First description of the 'Stochastic Diffusion Search' algorithm.]

[18] Bishop, J.M., Bushnell, M.J. & Westland, S., (1991), The Application of Neural Networks to Computer Recipe Prediction, *Color*, 16 (1), pp. 3-9. [Successful application of neural networks to color recipe prediction.

[19] Bishop, J.M., Keating, D.A. & Mitchell, R.J., (1998), A Compound Eye for a Simple Robotic Insect, in Austin, J., *RAM-Based Neural Networks*, pp. 166-174, World Scientific.[The use of weightless neural network in mobile robot localisation.]

[20] Bishop, J.M., (2002), Dancing with Pixies: strong artificial intelligence and panpyschism. in: Preston, J. & Bishop, J.M., (eds), (2002), *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*, Oxford University Press, Oxford UK. [First publication of the Dancing with Pixies (DwP) reductio against machine consciousness.]

[21] Bishop, J M, (2004), A view inside the Chinese room, *Philosopher* 28 (4), pp. 47–51. [Summary of Searle's Chinese room argument.]

[22] Bishop, J.M. (2009), A Cognitive Computation fallacy? Cognition, computations and panpsychism, *Cognitive Computation* 1(3), pp. 221-233. [Critique of the ubiquitous computational metaphor in cognitive science.]

[23] Bledsoe, W.W. & Browning, I., (1959), Pattern recognition and reading by machine, Proc. Eastern Joint Computer Conference, pp. 225-232. [First description of the n-tuple method for pattern classification later developed into the 'weightless neural network' paradigm by Igor Aleksander.]

[24] Bringsjord, S., & Xiao, H., (2000), A refutation of Penrose' Gödelian case against artificial intelligence, *J. Exp. Theo. Art. Int*. 12, pp. 307-329. *[Selmer Bringsjord's critique of Roger Penrose's Gödelian argument against artificial intelligence.]*

[25] Bringsjord, S., & Noel, R., (2002), Real Robots and the Missing Thought-Experiment in the Chinese Room Dialectic. in: Preston, J. & Bishop, J.M., (eds), (2002), *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*, Oxford University Press, Oxford UK. [Selmer Bringsjord's exposition of a 'missing thought experiment' in Searle's Chinese room argument].

[26] Broomhead, D. S. & Lowe, D., (1988), Radial basis functions, multi-variable functional interpolation and adaptive networks (Technical report). Royal Signals and Radar Establishment RSRE 4148, UK. [First exposition of the Radial Basis Function network.]

[27] Bryson, A.E., (1963), Optimal programming problems with inequality constraints. I: Necessary conditions for extremal solutions, J. AAAI, 1 (11), pp. 2544-2550. [Vapnik cites this paper by Bryson et al as the first publication of the 'back propagation' learning algorithm.]

[28] Carpenter, G. & Grossberg, S., (1987), A massively parallel architecture for a self organising neural pattern recognition machine, Computer Vision, Graphics and Image Processing: 37, pp. 54-115. [Early publication of the ART-1 classifier.]

[29] Chalmers, D., (1990), Why Fodor and Pylyshyn Were Wrong: The Simplest Refutation, In Proceedings of The Twelfth Annual Conference of the Cognitive Science Society, pp. 340-347. [David Chalmers' rebuttal of Fodor and Pylyshyn's critique of connectionism.]

[30] Chalmers, D., (1995), Minds, Machines, And Mathematics A Review of Shadows of the Mind by Roger Penrose, PSYCHE, 2: 9. [David Chalmers' critique of Roger Penrose's Gödelian argument against artificial intelligence.]

[31] Chalmers, D.J., (1996), *The Conscious Mind: In Search of a Fundamental Theory*, Oxford University Press, Oxford, UK. [David Chalmers' classic monograph espousing his 'psycho-functionalist' theory of mind.]

[32] Chrisley R., (1995), Why Everything Doesn't Realize Every Computation, Minds and Machines: *Minds and Machines Volume* 4, pp. 403-420.[Ron Chrisley addresses Hilary Putnam's critique of functionalism.]

[33] Chrisley R., (2006), Counterfactual computational vehicles of consciousness, Toward a Science of Consciousness (April 4th-8th 2006), Tucson Arizona, USA. [Ron Chrisley addresses Bishop's 'Dancing

with Pixies' reductio against machine consciousness.]

[34] Ciresan, D.C., Meier, U., Gambardella, L. M. Schmidhuber, J., (2010), Deep Big Simple Neural Nets For Handwritten Digit Recognition, *Neural Computation* 22 (12): pp. 3207?3220. [Deep learning networks applied to hand writing recognition.]

[35] Copeland B.J., (1997), The broad conception of computation, *American Behavioral Scientist* 40 (6), pp. 690- 716. [Copeland discusses foundations of computing and outlines so called 'real-valued' super-Turing machines.]

[36] Cortes, C. & Vapnik, V.N., (1995), Support-Vector Networks, *Machine Learning* 20 (3), pp. 273 - 297. *[Early discussion of the Support Vector machine*.]

[37] Deacon, T, (2012), *Incomplete Nature: How Mind Emerged from Matter*, W. W. Norton & Company. [All encompassing monograph emphasising the importance of appropriate embodiment for the emergence of mind.]

[38] Dinsmore, J., (1990), Symbolism and Connectionism: a difficult marriage, Technical Report TR 90-8, Southern University of Carbondale, USA. [The limitations of univariate knowledge representation in classical connectionist systems to arity zero predicates.]

[39] Dreyfus, H.L. & Dreyfus, S.E., (1988), Making a mind versus modeling the brain: artificial intelligence at a branch point, *Artificial Intelligence* 117: 1. [Historical review of the battle between connectionism and GOFAI in cognitive science.]

[40] Fodor, J.A., (1983), *Modularity of Mind: An Essay on Faculty Psychology*. The MIT Press.[In which Fodor postulates a vertical and modular psychological organisation underlying biologically coherent behaviors.]

[41] Fodor, J.A., Pylyshyn, Z.W., (1988), Connectionism and Cognitive Architecture: a critical analysis, *Cognition* 28, pp. 3-72. [The classic critique suggesting serious limitations of connectionist systems in comparison to symbol processing systems.]

[42] Freeman, A. (2003), Output still not really convinced, *The Times Higher* (April 11th 2003), UK. [Discussion of the Chinese room argument.]

[43] Freeman, A., (2004), The Chinese Room Comes of Age: a review of Preston & Bishop, *Journal of Consciousness Studies* 11 (5/6), pp. 156-158. [Discussion of the Chinese room argument and Preston/Bishop's edited text.]

[44] Fukushima, K, (1980), Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics 36 (4), pp. 193-202. [The Neocognitron: one of the first 'deep learning' connectionist systems.]

[45] Gardenfors, P, (2004), Conceptual Spaces as a Framework for Knowledge Representation, *Mind and Matter* 2 (2), pp. 9-27. [On the problem of modeling representations.]

[46] Garvey, J., (2003), A room with a view?, *The Philosophers Magazine* (3rd Quarter 2003), p. 61.[Discussion of the Chinese room argument and Preston/Bishop's edited text.]

[47] Gerstner, W. & Kistler, W., (2002), *Spiking Neuron Models: Single Neurons, Populations, Plasticity*. Cambridge University Press, UK. [Classic text on spiking neuron connectionist modeling.]

[48] Gibson, J.J., (1979), *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin. [Gibson's classical text outlining the radical 'ecological approach' to vision and visual perception.]

[49] Grossberg, S., (1976), Adaptive Pattern Classification and Universal recording: II. Feedback, expectation, olfaction, and illusions, *Bio. Cybernetics* 23: pp. 187-202. [Early exposition of the ART classifier].

[50] Harnad S., (1990), The Symbol Grounding Problem. *Physica* D 42, pp. 335–346. [Harnad's key reference on the classical 'symbol grounding' problem.]

[51] Harnish, S., (2001), *Minds, Brains, Computers: An Historical Introduction to the Foundations of Cognitive Science*, Wiley-Blackwell. [Excellent general introductory text to Cognitive Science.]

[52] Hebb, D.O. (1949). *The Organisation of Behaviour*. New York: Wiley & Sons. [The original

exposition of Hebbian learning.]

[53] Hinton, G, (2007), Learning multiple layers of representation, *Trends in Cognitive Science* 11 (10), pp. 428-434. [Hinton's discussion of deep learning.]

[54] Hodgkin, A. L. & Huxley, A. F., (1952), A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology* 117 (4), pp. 500?544. *[The classic paper on the detailed Hodgkin-Huxley model of the [squid] neuron.]*

[55] Hopfield, J..J., (1982), Neural networks and physical systems with emergent collective computational abilities, *Proc. Nat. Acad. Sci.* 79 (8), pp. 2554-2558. [Outlines the use of methods from physics to analyse properties of neural systems.]

[56] Hopfield, J. J. (1984) Neurons with graded response have collective computational properties like those of two-state neurons. *Proc. Nat. Acad. Sci.* 81, pp. 3088-3092. [The binary Hopfield neural network.]

[57] Hopfield, J. J. and Tank, D. W., (1985), Neural computation of decisions in optimization problems. *Biological Cybernetics* 55, pp. 141-146. [The use of Hopfield networks in optimization.]

[58] Hubel, D.H., Wiesel, T.N. (1962), Receptive fields, binocular interaction and functional architecture in the cat's visual cortex, *Journal of Physiology* 160 (1), pp, 106-154. [Hubel & Wiesel's classic paper on neuron behavior.]

[59] Irwin, G., Warwick, K. & Hunt, K., (eds), (1995), *Neural Network Applications In Control*. IET Digital Library, UK. [A selection of neural network applications.]

[60] Jaeger, H. & Haas, H. (2004), Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication, *Science* 304: 5667, pp. 78-80. [On why echo state networks learn very efficiently.]

[61] James, W, (1891), *Psychology (briefer course)*, Holt, New York, USA.[Early classic text on the principles of psychology.]

[62] Katevas, N., Sgouros, N.M., Tzafestas, S., Papakonstantinou G., Beattie, P., Bishop, J.M., Tsanakas, P & Koutsouris D., (1997), The Autonomous Mobile Robot SENARIO: A Sensor-Aided Intelligent Navigation for Powered Wheelchairs, *IEEE Trans. on Robotics and Automation* 4 (4), pp. 60-70. [The use of Stochastic Diffusion in mobile robot localization.]

[63] Kleene, S. C., (1956), Representation of events in nerve nets and finite automata. In Automata Studies 34, *Annals of Mathematics Studies*, pp. 3-42. Princeton University Press, Princeton, NJ. [Before Minsky, Kleene proved that rational-weighted recurrent neural networks with boolean activation functions are computationally equivalent to classical finite state automata.]

[64] Klein, C., (2004), Maudlin on Computation, (working paper).[On Maudlin's model of the implementation of computation.]

[65] Kohonen, T., (1988), *Self-Organising Maps*, Springer.[Introduction to Kohonen feature maps.]

[66] Kohonen, T, Barna, G. & Chrisley, R., (1988), Statistical Pattern Recognition with Neural Networks, In: Proceedings of the IEEE Conference on Neural Network 1: 61-68, 24-27 July 1988, San Diego, CA.[A discussion of practical issues relating to the Boltzmann machine and comparison of performance with other neural architectures.]

[67] Le Cun, Y., (1985), Une procedure d'apprentissage pour reseau a seuil a symmetrique (a Learning Scheme for Asymmetric Threshold Networks), Proceedings of Cognitiva 85, pp. 599-604, Paris, France. [Independent derivation of the back propagation learning algorithm.]

[68] Lighthill, J., (1973), Artificial Intelligence: A General Survey, in Artificial Intelligence: a paper symposium, Science Research Council, UK.[UK government-funded study providing a critique of GOFAI methods.]

[69] Lucas, J.R., (1961), Minds, Machines & Gödel, *Philosophy*, 36. [Lucas outlines the use of Gödelian argument against materialism.]

[70] Ludermir, T.B., de Carvalho, A., Braga, A.P. & de Souto, M.C.P., (1999), Weightless neural models: a review of current and past works, *Neural Computing Surveys* 2, pp. 41-61.[Recent review of the

weightless neural network paradigm.]

[71] Maass, W. & Bishop, C., (eds.), (1999), *Pulsed Neural Networks*, The MIT Press, Cambridge MA, USA. [Classic text on spiking neural networks.]

[72] Maass, W, Natschlager, T. & Markram, H., (2002), Real-time computing without stable states: A new framework for neural computation based on perturbations, *Neural Computation* 14 (11), pp. 2531-2560.[Learning rules for Liquid State machines.]

[73] Maass, W. & Markram, H., (2004), On the computational power of circuits of spiking neurons, *Journal of Computer and System Sciences* 69 (4), pp. 593?616.[On the use of the same connectionist hardware to compute different functions.]

[74] Maudlin, T., (1989), Computation and Consciousness, *Journal of Philosophy* 86, pp. 407-432. [On the execution of computer programs.]

[75] McCarthy, J., Minsky, M., Rochester, N. & Shannon, C., (1955), A Proposal for the Darmouth Summer Research Project on Artificial Intelligence. In: McCarthy, J., Minsky, M., Rochester, N. & Shannon, C., (2006) A Proposal for the Darmouth Summer Research Project on Artificial Intelligence August 31 1955, AI Magazine 27(4): 12-14.[Reprint of the proposal for the 1956 Darmouth Conference, along with the short autobiographical statements of the proposers.]

[76] McCorduck, P., (1979), *Machines who think*, Freeman & Co., NY. [An account of the Dartmouth Conference.]

[77] McCulloch, W.S., Pitts, W., (1943), A logical calculus immanent in nervous activity, *Bulletin of Mathematical Biophysics* 5, pp. 115-133. [On the computational power of the McCulloch-Pitts neuron model.]

[78] Minsky, M., (1954), Neural Nets and the brain model problem. PhD thesis, Princeton, USA. [Early work from Minsky on a connectionist theory of mind.]

[79] Minsky, M. L. (1967). *Computation: finite and infinite machines*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA. [After Kleene, Minsky proved that rational-weighted recurrent neural networks with boolean activation functions are computationally equivalent to classical finite state automata.]

[80] Minsky, M & Papert, S., (1969), *Perceptions: an introduction to computational geometry*, The MIT Press. [Devastatining critique of the power of single layer perceptrons.]

[81] Mumford, D., (1994), Neural Architectures for Pattern-theoretic Problems. In: Koch, Ch., Davies, J.L. (eds.), (1994), *Large Scale Neuronal Theories of the Brain*. The MIT Press, Cambridge MA. [In which Mumford observes that systems which depend on interaction between feedforward and feedback loops are quite distinct from models based on Marr's feedforward theory of vision.]

[82] Nasuto, S.J., Bishop, J.M., Lauria, S., (1998), Time Complexity Analysis of the Stochastic Diffusion Search, Neural Computation '98, Vienna, Austria. [Derivation of the computational time complexity of stochastic diffusion search.]

[83] Nasuto, S.J., Bishop, J.M.:, (1998), Neural Stochastic Diffusion Search Network - a theoretical solution to the binding problem, *Proc. ASSC*2, pp. 19-20, Bremen, Germany. [Examination of the binding problem through the lens of stochastic diffusion search.]

[84] Nasuto, S.J., Dautenhahn K., Bishop, J.M., (1999), Communication as an Emergent Metaphor for Neuronal Operation. In: Nehaniv, C., (ed), (1999), *Computation for Metaphors, Analogy, and Agents, Lecture Notes in Artificial Intelligence* 1562, pp. 365-379. Springer. [Outlines a new metaphor grounded upon stochastic diffusion search for neural processing.]

[85] Nasuto, S.J., Bishop, J.M., (1999), Convergence of the Stochastic Diffusion Search, *Parallel Algorithms* 14 (2), pp. 89-107. [Proof of the convergence of stochastic diffusion search to global optimum.]

[86] Nasuto, S.J., Bishop, J.M. & De Meyer (2009), Communicating neurons: a connectionist spiking neuron implementation of stochastic diffusion search, *Neurocomputing* 72: pp. 4-6, pp. 704-712. [Review of a communication metaphor - based upon stochastic diffusion search - for neural processing.]

[87] Newell, A., & Simon, H.A., (1976), Computer science as empirical inquiry: symbols and search,

Communications of the ACM 19 (3), pp. 113-126.[Outlines the Physical Symbol System Hypothesis.]

[88] Ng, A. & Dean, J., (2012), Building High-level Features Using Large Scale Unsupervised Learning, Proc. 29th Int. Conf. on Machine Learning, Edinburgh, Scotland, UK.[Deep learning neural network that successfully learned to recognize higher-level concepts (e.g. human faces and cats) from unlabeled video streams.]

[89] Newell, A., Shaw, J.C. & Simon, H.A., (1960), A variety of intelligent learning in a General Problem Solver, in Yovits, M.C & Cameron, S., (eds), '*Self Organising Systems*', Pergamon Press, NY. [Early discussion of the General Problem Solver.]

[90] Overill, J., (2004), Views into the Chinese Room: New Essays on Searle and Artificial Intelligence, *Journal of Logic and Computation* 14 (2), pp. 325-326. [Critique of the Chinese room argument.]

[91] Penrose, R., (1989), *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*, Oxford University Press, Oxford UK. [Roger Penrose's text criticising computational artificial intelligence from a Gödelian perspective.]

[92] Penrose, R., (1994) *Shadows of the Mind: A Search for the Missing Science of Consciousness*, Oxford University Press, Oxford UK. [Roger Penrose's second text criticising computational artificial intelligence from a Gödelian perspective.]

[93] Penrose, R., (1996) Beyond the Doubting of a Shadow A Reply to Commentaries on Shadows of the Mind, *PSYCHE*, 2 (23). [Roger Penrose replies to criticism of his Gödelian arguments.]

[94] Pinker, S. & Prince, A., (1988), On Language and Connectionism: Analysis of a Parallel Distributed Processing Model of Language Acquisition. In: Pinker, S., Mahler, J. (eds.), (1988), *Connections and Symbols*, The MIT Press, Cambridge MA.[Deficiencies in the representation of complex knowledge by classical neural networks.]

[95] Preston, J. & Bishop, J.M., (eds), (2002), *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*, Oxford University Press, Oxford UK. [Edited collection of essays on John Searle's Chinese room argument.]

[96] Psyche, Symposium on Roger Penrose's Shadows of the Mind, http://psyche.cs.monash.edu.au/psyche-index-v2.html, *PSYCHE* VOL 2. [Special issues on Roger Penrose's Gödelian arguments against AI.]

[97] Putnam, H., (1988), *Representation & Reality*, Bradford Books, Cambridge MA.[The appendix contains Putnam's refutation of functionalism.]

[98] Rapaport, W.J., (2006), Review of Preston, J. & Bishop, J.M., (eds), (2002), Views into the Chinese Room: New Essays on Searle and Artificial Intelligence, Oxford University Press, Oxford UK. In: A*ustralian Journal of Philosophy*, 94(1), pp. 129-145. [Review of work on the Chinese room argument.]

[99] Rey, G, (2002), Searle's Misunderstanding of Functionalism and Strong AI. In: Preston, J. & Bishop, J.M., (eds), (2002), *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*, Oxford University Press, Oxford UK. [Criticism of Searle's Chinese room argument.]

[100] Richeimer, J., (2004), Review of Preston, J. & Bishop, J.M., (eds), (2002), Views into the Chinese Room: New Essays on Searle and Artificial Intelligence, Oxford University Press, Oxford UK, *Philosophical Books* 45 (2), pp.162-167. [Review of work on the Chinese room argument.]

[101] Rosenblatt, F, (1958), Two theorems of statistical separability in the Perceptron. In: The Mechanisation of thought processes: proceedings of a symposium held at the National Physical Laboratory 1, pp. 419-449, HMSO London. [Early work by Frank Rosenblatt on the power of the Perceptron.]

[102] Rosenblatt, F, (1962), *The Principles of Neurodynamics*, Spartan Books, New York. [Introduction to Perceptrons.]

[103] Rumelhart, D.E., McClelland, J.L.., (1986), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, The MIT Press, Cambridge MA. [Seminal text which helped re-awaken interest in neural networks post Minsky & Papert's critique.]

[104] Rumelhart, D.E.., Hinton, G.E. & Williams, R.J., (1986), Learning representations by back-

propagating errors, *Nature* 323 (6088), pp. 533-536. [Independent derivation of the back-propagation algorithm.]

[105] Russel, B., (1905), *On Denoting, Mind* XIV(4), pp. 479-493, Oxford UK. [On quantified logical inference.]

[106] Schank, R C & Abelson, R P, (1977), *Scripts, plans, goals and understanding: An inquiry into human knowledge structures*, Psychology Press. [On the 'machine understanding' of stories.]

[107] Searle, J. R., (1980), *Minds, brains, and programs. Behavioral and Brain Sciences* 3 (3): 417-457, Cambridge, UK. [The Chinese room argument.]

[108] Searle, J., (1990), Is the Brain a Digital Computer?, *Proceedings of the American Philosophical Association* 64, pp.21-37. [Is computational syntax reducible to physics?].

[109] Searle, J., (1992), The Rediscovery of the Mind, pp. 227, MIT Press, Cambridge MA. [On computations and mind].

[110] Searle, J., (1994), The Mystery of Consciousness, Granta Books, London UK. [Interviews and reflections with a selection of twentieth century thinkers on consciousness.]

[111] Sejnowski, T.J., Rosenberg, C.R., (1987), Parallel networks that learn to pronounce English text. *Complex Systems* 1, pp. 145-168. [Example of an ANN which learns to pronounce English text.]

[112] Siegelmann, H. T., (1993), Neural and Super-Turing Computing, *Minds and Machines* 13, pp. 103-114.[On Siegelmann's claim that recurrent neural networks have super-Turing computational power.]

[113] Siegelmann, H. T., (1995), Computation Beyond the Turing Limit, *Science* 268 (5210), pp. 545-548. [On Siegelmann's claim that recurrent neural networks have super-Turing computational power.]

[114] Siegelmann, H.T & Sontag, E.F., (1984), Analog computation via neural networks, *Theoretical Computer Science* 131, pp. 331-360. [On Siegelmann's claim that recurrent neural networks have super-Turing computational power.]

[115] Smith P., (1998), *Explaining Chaos*. Cambridge University Press, Cambridge, UK. [Includes discussion of the 'shadowing theorems' in chaos].

[116] Smolensky, P., (1988), Connectionist Mental States: A Reply to Fodor and Pylyshyn, *Southern Journal of Philosophy* 26 (supplement), pp.137-161. [A refutation of Fodor and Pylyshyn's criticism of connectionism].

[117] Sprevak, M.D., (2005), The Chinese Carnival, *Studies in the History & Philosophy of Science* 36, pp. 203-209.[Reflections on Searle's Chinese room argument].

[118] Tanay, T., Bishop, J.M., Spencer, M.C., Roesch, E.B. & Nasuto, S.J., (2013), Stochastic Diffusion Search applied to Trees: a Swarm Intelligence heuristic performing Monte-Carlo Tree Search. In: Bishop, J.M & Erden Y.J, (2013), (eds), Proceedings of the AISB 2013: *Computing and Philosophy symposium,'What is computation?'*, Exeter, UK. [Describes the use of a swarm intelligence paradigm - stochastic diffusion search - to play the strategic game of HeX].

[119] Tesauro, G. 1989. Neurogammon wins Computer Olympiad. *Neural Computing* 1, pp. 321-323. [Playing backgammon using neural networks].

[120] Tarassenko, L., (1998), *A Guide to Neural Computing Applications*, Arnold, London. [Practical introduction to neural computing].

[121] Tassinari, R.P. & D'Ottaviano, I.M.L., (2007), Cogito ergo sum non machina! About Gödel's first incompleteness theorem and Turing machines, *CLE e-Prints*, Vol. 7 (3). [Discussion of Penrose arguments against artificial intelligence].

[122] Thompson, E, (2010), *Mind in Life*, Belknap Press. [Thompson's claim that mind is causally related to life].

[123] Torrance, S., Thin Phenomenality and Machine Consciousness, in R.Chrisley, R.Clowes and S.Torrance, (eds), (2005), Proc. 2005 Symposium on Next Generation Approaches to Machine Consciousness: Imagination, Development, Intersubjectivity and Embodiment, AISB05 Convention. Hertfordshire: University of Hertfordshire, pp. 59-66. [On differences between brain simulations and real

brains].

[124] Touzet, C., (1997), Neural Reinforcement Learning for Behavior Synthesis. in: N. Sharkey, N., (ed) (1997), Special issue on Robot Learning: The New Wave, *Journal of Robotics and Autonomous Systems* 22 (3/4), pp. 251–281. [In reinforcement learning Touzet details replacing the Q-table with an MLP network or Kohonen network.]

[125] Turing, A. M., (1937), [Delivered to the Society in November 1936], "On Computable Numbers, with an Application to the Entscheidungsproblem". *Proceedings of the London Mathematical Society*. 2 42. pp. 23-65, London, UK. [Introduction to the Turing machine and its application to the Entscheidungsproblem].

[126] Van Gelder, T., Port, R.F. (1995), It's About Time: an overview of the dynamical approach to cognition. In: Port, R.F. & Van Gelder, T. (eds.), (1995), *Mind as Motion*, pp. 1-45, The MIT Press, Cambridge MA. *[Introduction to the dynamical system theory of mind]*.

[127] Van de Velde, F., (1997), On the use of computation in modelling behaviour, Network: *Computation in Neural Systems* 8, pp. 1-32. [Critique of neural modeling in classical connectionist models of low level cognition].

[128] Varela, F.J., Thompson, E. & Roesch, E., (1991), *The Embodied Mind: Cognitive Science and Human Experience*. The MIT Press. [Varela et al's seminal introduction to the so called 'embodied concept' of mind].

[129] Waskan, Jonathan A., (2005), Review of Preston, J. & Bishop, J.M., (eds), (2002), *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*, Oxford University Press, Oxford UK. *Philosophical Review*, 114: 2, (2005). [Further reflections on the Chinese room argument].

[130] Werbos, P., (1974), Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences, PhD thesis, Harvard University. ['Usually' identified as the first derivation of the back propagation algorithm].

[131 Wheeler, M, (2002), Change in the Rules: Computers, Dynamical Systems, and Searle. In: Preston, J. & Bishop, J.M., (eds), (2002), *Views into the Chinese Room: New Essays on Searle and Artificial Intelligence*, Oxford University Press, Oxford UK. [In which Mike Wheeler uses the Chinese room argument to target all purely formal theories of mind].

[132] Whitehead A.N. & Russell, B., (1910), *Principia Mathematica* (3 volumes), Cambridge University Press, Cambridge UK. [An attempt to derive mathematics from 'first principles'].

[133] Wittgenstein, L., (1948), *Last Writings on the Philosophy of Psychology* (vol. 1). Blackwell, Oxford, UK. [The writings Wittgenstein composed during his stay in Dublin between October 1948 and March 1949, one of his most fruitful periods, which informed the Philosophical Investigations].

Biographical Sketch

Dr. Mark Bishop is *Professor of Cognitive Computing* at Goldsmiths, University of London; Director of the Goldsmiths centre for Radical Cognitive Science and Chair of the UK society for Artificial Intelligence and the Simulation of Behaviour (AISB). He has published over 140 articles in the field of Cognitive Computing: *its theory*, where his interests focus on the foundations of the swarm intelligence paradigm 'Stochastic Diffusion Search'; *its application*, where he has worked on industrial problems relating to autonomous robotics, neural networks and color; and *its philosophical foundations*, where he developed the 'Dancing with Pixies (DwP)' thought experiment, a novel *reductio ad absurdum* argument against the possibility of computational machine consciousness; he has also written extensively on John Searle's 'Chinese room' argument.