The Rise of Cognitive Science in the 20th Century

Carrie Figdor

(forthcoming in Kind, A., ed.,
Philosophy of Mind in the 20th and 21st Centuries: Routledge)
Penultimate Draft

Abstract.
This chapter describes the conceptual foundations of cognitive science during its establishment as a science in the 20th century. It is organized around the core ideas of individual agency as its basic explanans and information-processing as its basic explanandum. The latter consists of a package of ideas that provide a mathematico-engineering framework for the philosophical theory of materialism.

Keywords: cognitive science; psychology; linguistics; computer science; neuroscience; behaviorism; artificial intelligence; computationalism; connectionism

Part I. Introduction.

Cognitive science is the study of individual agency: its nature, scope, mechanisms, and patterns. It studies what agents are and how they function. This definition is modified from one provided by Bechtel, Abrahamson, and Graham (1998), where cognitive science is defined as "the multidisciplinary scientific study of cognition and its role in intelligent agency." Several points motivate the modification. First (and least consequential), the multidisciplinarity of cognitive science is an accident of academic history, not a fact about its subject matter (a point also pressed in Gardner 1985). Second, the label "intelligent" is often used as a term of normative assessment, when cognitive science is concerned with behavior by entities (including possibly groups, as individual or collective agents) that are not considered intelligent, as well as unintelligent behavior of intelligent agents, for any intuitive definition of "intelligent".[1]

Third and most importantly, the term "cognition" is omitted from the definiens to help emphasize a position of neutrality on a number of contemporary debates. Cognition can often reasonably be equated with mental activity, but the mind has traditionally been associated or contrasted with the brain. The modified definition recognizes that whether or how much cognition is brain-based is a matter of considerable dispute (e.g., Clark 1997; Gallagher 2005; Adams and Aizawa 2008; Chemero 2011; Kiverstein and Miller 2015). That said, for reasons of brevity of exposition I will often write in terms appropriate to the traditional brain-based framing of cognition.

In addition, the scope of cognition (and agency) is currently in flux. For example, if plants have cognition (Trewavas 2005; Calvo and Keijzer 2009), then brains and animal bodies are not required for cognition or agency. Other writers are more restrictive. For example, Von Eckardt 2003 sees the domain of cognitive science as the *human*

[1] For example, Newell and Simon's (1976) physical symbol system hypothesis – that a physical symbol system has the necessary and sufficient means for intelligent action – covered humans and computers alike. They agreed that only systems of sufficient complexity and power could exhibit general intelligence, but intelligent action was not necessarily human action.

cognitive capacities. My working assumption is that human-style cognition is a special, if prototypical, case. Many people are most interested in human cognition. But what counts as cognitive will ultimately depend on the systems to which the basic conceptual framework of cognitive science can be fruitfully applied.

As conceptual history, the rise of cognitive science is the story of the articulation of the core concepts for explaining agency. This article explains five key innovations comprising the *basic explanatory package* of cognitive science. In traditional philosophical terms, they constitute the conceptual framework for explaining how the mind could be material. This package unifies the field despite the remaining conceptual and practical impediments of disciplinary boundaries, internal debates about how the package should be refined, and its incompleteness. It is assumed here that this package will be elaborated, not abandoned, in future work, just as the theory of evolution has continually unified biology despite tensions and controversies about its proper form.

The foundational ideas are associated with the main contributors and their main works; regrettably, discussion of the contributions of historical precursors and important contemporary figures is omitted for space reasons.[2] These ideas include the information-processing program (Alan Turing), neurons as information-processors (Warren McCulloch and Walter Pitts), feedback control of processing (Norbert Wiener), information as a measure of the structure of communication (Claude Shannon), and information-processing as a multiperspectival explanatory framework (David Marr). These ideas can be briefly described as follows, with details provided below.

Turing showed how recognizably rational behavior could be produced by an agent if very few distinct types of simple internal state transitions were sequenced in the right way. A human computer added columns of numbers, and so too could a simple Turing machine sequenced in the right way. McCulloch and Pitts showed how the basic internal machinery of the brain could be seen to realize these rational transitions. They mapped inferential steps involving propositions to transitions in states of neurons. Wiener described how the future behavior of such agents could depend on the impact of their prior responses on their environment. An agent can learn from experience when it can adaptively modify its behavior in response to experience that is itself a consequence of its prior behavior. Shannon showed how information could be understood and quantified in terms of the statistical or probabilistic structure inherent in communication. This structure is derived from conventional regularities that agents jointly create and can individually exploit to help achieve their goals. Marr showed that information-processing explanations shared an explanatory structure in which goals, processing steps, and physical operations would all be specified. This explanatory framework applied to non-rational as well as rational processes.

Of the five, Turing's and Shannon's contributions may be most fundamental: they articulated the core concepts of "processing" and "information" in "information-processing". In the case of Shannon, there are many other technical as well as colloquial concepts of information (Adriaans 2012). The claim made here is that Shannon's concept is basic to cognitive science, and that its explanatory potential (unlike that of Turing's

[2] Besides Bechtel et al. op.cit., Boden 2006 is an authoritative and comprehensive discussion. Apray 1985: 120 provides a detailed chronology of key relevant works.

model) has barely begun to be elaborated. I discuss its relation to the philosophical notion of representation or content in Section II.4.

A potential sixth core element of the package is a theory of the goals of and constraints on information-processing capacities at the agent level. Proposals for the social root of intelligence (e.g., Jolly 1967; Humphreys 1976; Dunbar 1998; Sterelny 2007) are attempts to make theoretical sense of agents' goals, assessments, expectations, and responses within their social contexts. Developing and integrating a basic framework of agentic goals vis-à-vis other agents is one of the main challenges facing 21st century cognitive science.

The discussion below emphasizes the abstract nature of the core ideas. This feature has led, I believe, to some misunderstanding about the relation between cognitive science and the discipline-specific ways in which the ideas were initially appropriated, articulated, investigated, and deployed in explanation. For example, Turing's model did not come with fine print stating the limits of its explanatory power. As we are discovering, much can be done in artificial intelligence to satisfy military, industrial, and commercial aims without addressing the symbol grounding problem – the problem of fixing the reference of symbols or concepts. Solving this problem may be crucial for explaining some aspects of agency, but Turing's bare-bones model is not sufficient to solve it. That is why it is just a part of the basic explanatory package.

Similarly, the fact that post-behaviorist empirical psychology proceeded without looking at the brain is not the denial of an essential explanatory connection in cognitive science (nor, for that matter, in psychology). Scientific investigation involving the brain had to wait until the 1990's. That was when the technology to measure ongoing neural activity with some degree of specificity during the performance of cognitive tasks became widely available. So when Searle (1980: 421**)** stated that "the whole idea of strong AI is that we don't need to know how the brain works to know how the mind works … [W]e can understand the mind without doing neurophysiology," this may be true of strong AI and parts of psychology yet false of cognitive science. The cognitive science-biology boundary is not yet fixed.

Finally, while the core ideas are abstract, they are fundamentally mathematical rather than philosophical, quantitative rather than qualitative. The genius of those contributing to the package was their ability to build conceptual bridges between intuitive conceptions of mind and non-intuition-based explanations of them. Philosophers have contributed significantly to cognitive science from the start – as critics (e.g., Searle 1980, Dreyfus 1992), integrators (e.g., Fodor 1983), collaborators (e.g., Churchland and Sejnowski 1992), champions (e.g., P.M. Churchland 1990, P.S. Churchland 1986), and theoreticians (e.g., Fodor 1975, Dennett 1987; Chalmers 1995). They will continue to do so not just in one or more of these roles (e.g. Block 2007), but also as disseminators (Hohwy 2014), participants (Eliasmith 2013), and articulators of new social and moral concerns that arise as intuitions about human cognition and agency are challenged (Roskies 2010; Allen, Varner, and Zinser 2000). We think about the mind differently now than we did 100 years ago, due to both theoretical and empirical advances. Future philosophical participation in cognitive science will have to take this change into account.

Part II. The Basic Explanatory Package

We do resent the hiatus between our mental terminology and our physical terminology. It is being attacked in a very realistic fashion today.

McCulloch 1943 (from the Warren S. McCulloch Papers, cited in Piccinini 2004)

Cognitive science aims to explain agency in material terms – in particular, in mathematical terms that bridge logic (mind) and engineering (matter). Oddly, mathematics is omitted from the list of disciplines contributing to cognitive science even though many pioneers of cognitive science, including Turing, Pitts, Wiener, and Shannon, were mathematicians. In contrast, neuroscience, philosophy, psychology, linguistics, and computer science are usually listed as constitutive disciplines (e.g., Bechtel et al. op.cit.: 69-70; Miller 2003: 143; Heckathorn 1989) even though (like mathematics) most areas of these disciplines have nothing to do with cognitive science. Anthropology is also included even though it quickly parted ways from cognitive science (Bender et al. 2010, 2013). Sociology or "sociocultural studies" (Bechtel et al. op.cit.: 93) is mainly noted for its absence (Bainbridge 1994: 408), underlining the lag in integrating social aspects of cognition.

The omission of mathematics may be due to the fact that until Turing we lacked an empirically plausible model of how the mind could be material. Without such a model materialists could do little to counter the intuition, and philosophical position, that the mind is exempt from the mathematico-engineering, mechanical explanation of the rest of nature. Gottfried Leibniz (a mathematician) had the idea of a logical calculus in the 17th century, but he also denied that perception and consciousness could be implemented in a machine (Monadology 17).[3] With Turing's breakthrough, we could retrospectively identify percursors – more mathematicians. In the 18th century, Charles Babbage invented (but did not fully build) an analytical engine for general computing that operated on the same principles as the Jacquard loom, which used sequences of punchcards to organize sequences of the machine's weaving operations (Copeland 2008). George Boole (1854) found that mathematical operations performed on sets could also be logical operators that operated on propositions or sentential thought contents, suggesting that the resulting operations were laws of thought. Gottlob Frege (1879) added a logic that allowed for operations on parts of propositions, formalizing deductive inference.

The study of these ideas, blended in mathematical logic, unified the conceptual founders of cognitive science (Aspray 1985). The ideas themselves provided materialists with a clear engineering target: to build something that can do these logical operations.

[3] The influence of Leibniz's logic on 19th century logicians is disputed (Peckhaus 2009), although Wiener (1961: 12) calls Leibniz the "patron saint" of cybernetics and Shannon (1948: 52) in turn credits Wiener as an important influence. What is indisputable is that the isolated idea of a logical calculus had no impact on the development of a materialist alternative to dualism prior to Turing, who relied directly on Boole, as did Shannon; meanwhile, Pitts was a student of Carnap, and Newell, Shaw, and Simon demonstrated the information-processing paradigm's possibilities when their Logic Theorist program provided a more elegant proof of a theorem from Russell and Whitehead's Principia Mathematica than the one in Principia (which led them to try, without success, to publish this result in a paper that listed Logic Theorist as a co-author).

II.1. 1936: Turing: Software

Turing (1936) provided the first explanatory link between these logical operations and a machine that could perform them. He showed that any well-defined logical or mathematical problem that had an effective solution – that could be solved in a finite number of steps – could be solved by following simple state transitions in a sequence. Although a Turing machine was not a physical device, each step could be imagined physically as a series of squares on a tape plus a read-write device. The device would scan a square (start the transition), erase or print a 1 or 0 on the square (perform a simple operation), and move to the next square (end the transition). These state transitions could be realized by a physical device with appropriate on/off switches as 1's and 0's and a way to distinguish and respond to them. 1's and 0's are numerically, not psychologically, interpreted states, but the way was open to interpret thoughts as complex symbols that could be similarly manipulated. Turing also showed that given enough space and time a single sequence of simple steps – a universal Turing machine – could encompass any other sequence by embedding them (or inserting them as needed) in the larger sequence. Like a mind, a universal Turing machine was versatile ("general-purpose"): it could solve "any problem that can be reduced to a programme of elementary instructions" (Williams and Kilburn 1948).

But *can* all mental operations be reduced to a series of elementary instructions? Descartes argued that animals lacked minds because they lacked language, the means by which humans can express an infinite variety of thoughts. (He did not consider prelinguistic infants, inter alia.) But whether universal Turing machines were as versatile as minds did not have to turn on intuitive measures of versatility. As Alan Newell, Cliff Shaw, and Herbert Simon – pioneers in developing computer programs with psychologically interpretable states and transitions[4] – put it:

> [A] program incorporating such [elementary information] processes, with appropriate organization, can in fact solve problems. This aspect of problem solving has been thought to be "mysterious" and unexplained because it was not understood how sequences of simple processes could account for the successful solution of complex problems. The theory dissolves the mystery by showing that nothing more need be added to the constitution of a successful problem solver. (1958: 152)

"Dissolves" may be overstating matters, but the demystifying of mind had begun.

Turing's theory left open how an embodied universal Turing machine might be designed. The first programmable computers, which were built in the 1940's (Williams and Kilburn op.cit.; von Neumann 1945; Godfrey and Hendry 1993), were designed to meet engineering goals. For example, optimizing operational efficiency by means of central program-storage unit (a Central Control) entailed minimizing the flexibility of the operations (von Neumann 1966: secs. 2.2, 2.3). But so what, if any needed flexibility could be left up to a human programmer? Similarly, ease of repair could be optimized by

[4] Newell, Shaw, and Simon (1958) developed the first list-processing language (IPL) for an information-processing system of psychologically interpretable transitions, rather than transitions in terms of 1's and 0's (Boden 1991: 10).

building a “fragile” machine that would stop operating at an error (von Neumann 1966: 73), even if this meant they did not operate like brains, which isolated problems for working on them on the side.

Such engineering decisions should not be confused with limits on the explanatory potential of Turing’s model. Dreyfus (1992) argued that computing can’t explain human intelligence because the latter is context-sensitive and thus not rule-governed. Similarly, the fragility of von Neumann-style computers was treated as a bug by early champions of connectionism (e.g., Churchland 1990), a computing design based on neurophysiology (described below). Fragility, flexibility, and context-dependence are concepts in the same family as the intuitive idea of versatility. Turing’s model left open how any of these features might be realized in a universal Turing machine, and is consistent with a continuum of cognitive systems or agents of different degrees of versatility.

Nevertheless, the immediate assimilation of minds to computers by some psychologists and early AI researchers revealed exuberant hopes for how much mind could be explained with these first incarnations of Turing’s model. Due to arguments showing that no formal logical system could be used to prove all formulas that we recognize as being true, Turing was aware that a simple Turing machine could not do everything a human mind could do (Copeland and Shagrir 2013). But Turing (1950) also linked his processing story to human linguistic behavior by proposing the Turing Test, in which an interrogator tries to determine if her hidden interlocutor is a human or a computer. He predicted a computer would pass the Turing Test within 50 years; it remains unpassed. Simon reportedly predicted in 1957 that a computer would beat a human chess champion within 10 years; Big Blue beat Gary Kasparov in 1997. Searle’s (1980) Chinese-room thought experiment, which concludes that there is no understanding in a system that realizes an unelaborated Turing machine, provided a sharp rhetorical counter to these claims.

The early exuberance may also have reflected the fact that to experimental psychologists Turing’s model provided a viable non-introspectivist alternative research programme to behaviorism. In the early days of scientific psychology, introspectivist or structuralist psychologists (such as Wilhelm Wundt and Edward Titchener) used the reports of trained introspectors as evidence for the workings of the mind. When introspectors disagreed, there was no objective criterion for determining who might be right. Such unresolvable conflicts discredited structuralism as scientific psychology. Radical behaviorism went to the opposite extreme: the only allowable evidence was observable behavior or environmental contingencies, and only behavior needed to be explained. Behaviorism in this radical form never took hold in developmental, comparative, social, perceptual, or clinical psychology, and was not dominant outside the U.S. (Greenwood 1999; Miller 2003); even B.F. Skinner, its most well-known defender, was conflicted about it (Baars 2003). But where it was influential, its influence was profound: Neisser’s 1967 *Cognitive Psychology*, hailed as the ur-text of post-behavioristic experimental psychology, had six chapters on vision, four on audition, and just one slim final chapter on higher cognition.[5]

[5] Radical behaviorism did leave two important legacies. First, the demand for observable behavioral evidence of psychological claims (“methodological” behaviorism) is now entrenched. Second, by focusing on behavior rather than consciousness, behaviorism

As stored-program computing took off, experimental psychologists were facing a growing pile of anomalies that motivated looking inside the behaviorist's black box. Miller (1956) showed that short-term memory capacity stayed constant at around 7 'chunks' of information because items could be recoded into new 'chunks': for example, a 10-digit number is more easily remembered by being recoded into 3 chunks (e.g., 123-456-7890). This showed that internal cognitive machinery was needed to explain memory. Chomsky (1959) argued that children's linguistic output was governed by grammatical rules (or violated those rules in regular ways) that were underdetermined by the speech they heard as stimulus. This evidence of the 'poverty of the stimulus' (its inadequacy to explain the output) showed that internal operations were needed to explain language.

These and other results made the emerging cognitive science of information-processing highly attractive: it seemed "complicated enough to do everything that cognitive theorists have been talking about" (Miller, Galanter, and Pribram 1960: 43). What they had been talking about, inter alia, were ways to explain phenomena that made behaviorism implausible. Thus, psychologists took away from Turing the lesson that "if they could describe exactly and unambiguously anything that a living organism did, then a computing machine could be built that could exhibit the same behavior with sufficient exactitude to confuse the observer" (Baddeley 1994: 46).

No wonder, then, that Turing's model was immediately elaborated at a level appropriate to human-centered psychology: the symbols were interpreted as natural-language-like concepts or mental representations, and the rules were the rules of deductive logic or heuristics (Fodor 1975; Newell and Simon 1976; Miller, Galanter, and Pribram op.cit.: 3). This 'rules-and-representations' research programme came to be known as classical computationalism. The stored-program computer of von Neumann's design was the machine for which these first psychologically-interpreted internal state transitions were developed. They were specified in the form of software programs written in high-level programming languages.

The autonomy of psychology from biology (or neuroscience in particular) should also be understood in this context. Off-loading problems that are not of direct interest, particularly if the technology for investigating them is not yet available, is a rational scientific strategy. As Newell, Shaw, and Simon (op.cit.: 163) put it: "Discovering what neural mechanisms realize these information-processing functions in the human brain is a task for another level of theory construction." The Turing-inspired research left open how much progress could be made without engaging with other levels of theory construction.

II. 2. 1943: McCulloch and Pitts: Brainware

A materialist explanation of agency requires a theory of how physical agents could be cognitive systems. Assuming humans as the prototype of such an agent, McCulloch (a neurophysiologist) and Pitts (a mathematician) provided this theory. They proposed that neurons were biological logic gates.

A logic gate is a unit whose operations can be interpreted in terms of the truth table for the logical operations of 'and' and 'or', the operations in Boolean logic. An 'and' gate fires a pulse if and only if its two input channels both fire, mirroring the way a

"helped to break down the distinction between the mental behavior of humans and the information processing of lower animals and machines" (Aspray (1985: 128).

conjunction – A and B – is true if and only if both A and B are true. An 'or' gate fires if at least one of its two input channels fires, mirroring the way a disjunction – A or B – is true if and only if at least one of the constituent sentences is true. A McCulloch-Pitts neuron is an abstract biological analogue of an electrical switch or relay, a basic component of a von Neumann computer (von Neumann 1945: 4.2, 4.3; Wiener 1948: Ch. 5; Arbib 2000: 212). McCulloch-Pitts neurons were binary in operation, so their states could be associated with propositions: activation could be associated with truth values (on/1/true, off/0/false) and patterns of activation with inference. While such sparse coding (i.e., 1 activated neuron = 1 true proposition) is empirically wildly implausible, this interpretation is the simplest that directly links Turing's model, with its simple state transitions, to the activity of the basic operating units of actual brains.[6] This link presupposed the discovery by neuroscientist Santiago Ramon y Cajal that neurons do not form a continuous net but are discrete units that stand in electrochemical relations.

The McCulloch-Pitts theory inspired connectionist or neural network computing. Connectionist networks are virtual collections of McCulloch-Pitts neurons running on standard computers. They have simple units (nodes) with connections to other nodes. Input nodes are analogous to sensory neurons, output nodes to motor neurons, and "hidden" layers of nodes to neurons that mediate between input and output. Numerical weights on the connections regulate the amount of input (activation) passed or propagated from one node to another. When a node obtains sufficient net input from its incoming connections to reach or pass a firing threshold, it sends input (fires) to the nodes to which it is connected by its outgoing connections. The weights on the connections at one stage of processing determine the activation pattern at the next stage.[7]

Connection weights implicitly contain the record of past activation and so collectively embody what the network has learned from experience. The weights are adjusted automatically or by a human modeler using a learning rule. For example, a simple Hebbian learning rule (after psychologist Donald O. Hebb) increases the numerical value assigned to the connection between two nodes that co-activate. This makes them more likely to be co-activated in the future, mimicking the neurophysiological feature that synaptic connections are strengthened when two neurons are co-activated (called long-term potentiation, or, as the slogan goes, "neurons that fire together wire together").

Connectionist-style modeling of cognitive capacities began in the 1940's and 1950's but was overshadowed by programming research until the 1986 publication of *Parallel Distributed Processing* (Rumelhart, McClelland, and the PDP Research Group),

[6] Von Neumann suggested a further analogy: the Central Control and Memory of a standard stored-program computer were intended to "correspond to the associative neurons in the human nervous system" (von Neumann: 3, sec. 2.6; sec. 4.0, 4.2) – that is, the hidden layers of a connectionist network.

[7] This description of neural networks best fits feedforward networks, such as those in the PDP Research Group papers cited below. In these networks, activation passes from input to hidden to output layers, and the output is what the nodes in the output layer compute. Another important strand of connectionism stems from Hopfield (1982), who designed a recurrent network. In a recurrent network, every node provides input to every other node, and the network's output is a stable activation pattern of the whole network.

which gathered papers on neural net research in perception, verb parsing, and other capacities. However, while early champions of connectionism approvingly contrasted their brain-like architecture with that of stored-program computing, McCulloch-Pitts neurons are no less abstract than the squares on a Turing machine tape. For example, there is no distinction between kinds of neurons and no means to represent the role of neuromodulators in realizing the context-dependence and variability of neural signaling (Dayan 2012; Izhikevich 2007). In fact connectionist networks are now used to model all kinds of networks (Baronchelli et al. 2013). Nodes and weighted connections (now also called edges) can represent, respectively, agents and the relative importance of interagent relations (Froese 2014); words and their frequency of association (Borge-Holthoefer and Arenas 2010); and ideas and the spread of innovation (Mason et al. 2008).

That said, there are important differences. In classical computing there is one series of computations,, represented by symbolic-program-governed operations on squares of tape. (More than one series can be run in parallel, but they are equivalent to a single series.) In a connectionist network, multiple computations – each represented by the equation-governed activation of each neural logic gate – go on simultaneously. In classical computationalism the problem is to write a program that will generate the desired output given the input; in connectionism the problem is to get the connection weights set so that the desired output is generated from the input. These differences yield interesting differences in terms of their explanatory power. Serial, stored-program computing is terrific for modeling logical operations, while parallel, weighted-connection computing is terrific for partitioning data into classes by frequency of association.

The relations between these types of computing and between each type and psychological processes are still debated. One way this debate has been framed is whether connectionist networks describe a cognitive level directly or whether they implement classical computation (Fodor and Pylyshyn 1988; Smolensky 1991; Marcus 2001; Aizawa 2014). Currently, the activation patterns of the hidden layers in neural networks that are used to model brain activity have no clear psychological interpretation. Whether these patterns need to be so interpreted is also a matter of debate (Ramsey 2007; Bechtel 2001).

II. 3. 1948: Wiener: Feedback Control

Turing's model did not say how symbols or rules for manipulating them could be modified. Since many agents learn from experience, their agency cannot be explained by Turing machines that lack an internal learning mechanism. Wiener provided a model of feedback control, building on ideas from 19th century physicist James Clerk Maxwell.

Wiener (with his collaborator physiologist Arturo Rosenblueth) coined the term "cybernetics" (from the Greek for "steersman", 1961: 11) for the study of "control and communication in the animal and the machine" – physical systems, living or not. A feedback loop is an agent-environment causal loop (or an epicycle in it) that allows for adjustment of the agent's behavior (or a stage of it) in the light of what occurs in the environment as a result of its prior behavior. To use Wiener's example (1961: 7), the muscle motions involved in picking up a pencil require some sort of information that will guide the appropriate motor commands at each moment in a way that depends on how much farther away the pencil is at any moment. The motion of your arm, hand, and fingers at any time depends on the way the environment now affects your eyes (the source of the visual input of your arm position relative to the pencil) which depends on the motion you made a moment ago.

Cybernetics complicated the core explanatory package structurally and conceptually. In a simple Turing machine, the dependency between two states is set by a rule. Providing the initial input is like tapping the first domino in a series. In a simple feedforward neural network – in which connections propagate activation in one direction, from input to output – activation in nodes closer to output nodes cannot affect activation in nodes closer to input nodes. The updating of the network's connection weights by the network modeler is analogous to thought-insertion. In both cases, internal feedback loops are needed to enable outputs at a later stage to be used as input in an earlier stage. Of course, a system may be able to get feedback but not be able to use it to alter its behavior. Where there is feedback control, there is also the capacity to change behavior by using feedback. Where in addition the change in behavior is adaptive, or responsive to environmental contingencies, there is also learning.

In this way cybernetics also introduced the concepts of goals, expectations, and assessments into the basic explanatory package: a system that has the capacity to generate and use feedback to control its behavior adaptively is a system with goals (or final states), expectations (intermediate states), and ways to assess its input in the light of these expectations and goals. The feedback control concept applies to "a learning system that *wants* something, that adapts its behavior in order to maximize a special signal from the environment" (Sutton and Barto 1998: Preface). Understanding such a system requires understanding the many ways in which it is coupled with its environment.

Like the other elements of the core explanatory package, the cybernetic model is abstract enough to apply to a wide range of systems. Like them, too, cybernetics was elaborated early on in psychological terms. Miller, Galanter, and Pribram (1960) adopted the model to describe "how actions are controlled by an organism's internal representation of its universe." Their motivation was clear:

> The men who have pioneered in this area [of computing and programming] have been remarkably innocent about psychology – the creatures whose behavior they want to simulate often seem more like a mathematician's dream than like living animals. (op.cit.: 3)

They theorized that stimulus and response were stages of the same complex feedback loop, which they called a TOTE unit ("Test-Operate-Test-Exit"). What an organism did was guided by the outcomes of TOTE units, which could be organized hierarchically (that is, feedback loops within feedback loops). Such complications were critical for the information-processing paradigm to even begin to explain human agency.

More recently, the cybernetic idea is reflected in the predictive error minimization or Bayesian brain model of whole-brain function (Friston 2010; Clark 2013; Hohwy 2014), presaged by Rosenblueth, Wiener, and Bigelow (1943). A Bayesian model is one in which a system's states (often interpreted as its beliefs or hypotheses) are updated using Bayes' theorem. The theorem calculates the adjustments in the level of belief or credence one should have in a hypothesis in the light of new evidence and one's prior credence in that hypothesis. On the predictive brain model, the brain (or a structure within it) compares a new input value to an expected value, calculates the difference or error, if any, between the expected value and the actual input value, and makes an adjustment so that at the next stage its subsequent input is closer to its expectation. The system can adjust the hypothesis that generated the initial expected value to get a new expectation and then act much as it did, or it can adjust its subsequent behavior to get new input that will more closely match its expected value, or a bit of both.

When a feedback control loop is spatiotemporally tight, it is tempting to argue that a system does not require internal models or representations to explain its behavior. To borrow van Gelder's (1995) illustrative example, the Watt governor for a steam engine continuously and mutually adjusts linear motion and centrifugal force because these forces are realized by mechanically coupled parts (a throttle valve, a spinning spindle with weighted arms). But not all feedback control loops are so tight or so closely linked to sensorimotor capacities (as with Weiner's own example of reaching for a pencil). For example, reinforcement learning, when rewarded behavior becomes more frequent, falls squarely within the cybernetic model and yet requires non-behavioristic explanation (Rescorla 1988). As the predictive brain hypothesis is critically examined, the debate over the need for representational notions in neural networks (and, by implication, brains) is likely to expand to include the concepts of goals, expectations, and assessments that are integral to cybernetics.

II. 4. 1948: Shannon: Information

So far the explanatory package has focused on the "processing" in "information-processing". But what is information? Shannon's (1948) answer, building on Nyquist (1924) and Hartley (1928), is derived from his theory of communication. Communication is information transfer between agents. A core concept of information can be extracted from agents' coordinated communicative actions, which can be quantified.

Warren Weaver, Shannon's collaborator and communicator, distinguishes three basic problems in communication: the technical problem of accurate transfer of information from sender to receiver (was the message transmitted accurately?); the semantic problem of interpretation of meaning by the receiver as compared to the intended meaning of the sender (was the message understood in the intended way?); and the effectiveness problem of the success with which the meaning conveyed to the receiver leads to the receiver's desired conduct (did the message lead the receiver to respond as the sender intended?). Answers to the latter two questions are constrained by answers to the first. As Shannon (1948: 1) notes, the semantic aspect of communication, and specifically the problem of reference, is irrelevant to "the engineering problem" of information transfer. It does not follow that his solution to the engineering problem is irrelevant to explaining reference or intentionality – that is, the ability of minds to represent aspects of items in the external world, in such a way that it is also possible for them to *mis*represent). To the contrary, the theory describes the characteristics of a communication system that make reference possible. The link from reference to intentional contents – paradigmatically, thoughts about objects and their properties – will depend on how language and thought are related. For brevity, I assume here that at least some contentful mental states are partly, if not determinately, encoded in brain states (Dennett 1975), and that language expresses these contents, however imperfectly.

Shannon's theory "is specifically adapted to handle one of the most significant but difficult aspects of meaning, namely the influence of context" (Weaver 1949):

> The concept of information applies not to the individual message, as the concept of meaning would, but rather to the situation as a whole, the unit information indicating that in this situation one has an amount of freedom of choice, in selecting a message, which it is convenient to regard as a standard or unit amount.

In philosophy, a linguistic context is typically an extra-linguistic setting in which an utterance occurs, described in qualitative terms – who is talking to whom, when, where, about what. Here, a linguistic context is the structure of the language to which the message belongs and which constrains the meanings that can be communicated. This linguistic context is quantified in the theory, and a quantitative concept of information can be extracted from it. As Weaver (op.cit.: 11) puts it, "information relates not so much to what you do say, but to what you could say." Shannon's theory, like the other elements of the core explanatory package, is apt for many kinds of agents and communication systems, such as neural signalling (Dayan and Abbott 2004; Dayan and Abbott 2001). But for brevity I focus on the primary case of human linguistic communication.

In human language, the basic constraints on the set of possible messages is given by the statistical structure of the language in which source and receiver participate. The statistical structure of a language is reflected in its written form, which encodes the spoken form that directly expresses thought. The first letter of a sentence is maximally uncertain; it is most informative (has the most information) in that it constrains all subsequent letter choices while the only constraint on it is that the language contains that letter.[8] The frequencies of and relationships between letters can be quantified. The more the first choice constrains the second, the less information the second letter will contain: if the first letter is "Q" then given the features of English it is overwhelmingly likely that the next letter will be "U". English is about 50% redundant (for strings of up to 8 letters): about half the letters or words we use are chosen by us, and about half are determined by the statistical structure of English. This is why we can figure out badly garbled or incomplete messages.[9]

In short, in communication the U is redundant; it contains no more information than was given by the choice of Q; its presence is far from random; it is highly probable given the selection of Q; its entropy is low; we experience no surprise upon seeing a U; once you see a Q you already know what comes next. These are all ways of expressing the same probabilistic relationship that is the basis of the unit of information. A unit of information is a measure of how much freedom a source has in selecting a message. Information transfer can be quantified in terms of the probabilities assigned to each message in a set of possible messages that a sender could select to send to the receiver. The larger the set of possible messages, the more source freedom; the more source freedom, the more receiver uncertainty regarding which message will be selected. Greater

[8] In languages with non-alphabetic scripts (e.g. Chinese), the set of conventions behind the statistical structure of communication (discussed below) are presumably divided up differently from the way they are in alphabetic languages.

[9] At http://karpathy.github.io/2015/05/21/rnn-effectiveness/ the text that the network modeler's system generates illustrates the way that the statistical structure of English constrains letters to the extent that meaningful text emerges.

freedom of choice, greater uncertainty, and greater quantity of information go hand in hand (Weaver 1949).

What is not stated explicitly in Shannon's theory is the fact that the statistical structure captured in letter frequencies encodes some (but not all) of the conventions that create a language, distinguishing utterances or inscriptions from noise. Meaningfulness involves further conventions, which have not yet been modeled quantitatively. The place to look for these might well be in anthropology (Bender et al. 2010) and other cultural and social sciences. In philosophy, we have qualitative theories that focus, in philosophy of mind, on agents' interactions with enduring objects (Dretske 1988; Millikan 1985) and, in philosophy of language, agents' interactions with each other (Grice 1957; Lewis 1969). The concept of information in standard informational theories of content (e.g. Dretske op.cit.) is in effect pure reference, divorced from and independent of communication. Shannon's theory prompts thinking of reference as the upshot of additional constraints on communication, while leaving open how constrained a communication system must be, and which constraints it must have, in order for agents using that system to count as having representations (or intentionality) in the philosophical sense.[10]

In this vein, Weaver (op.cit.:14) speculatively adds into the communication process a step of statistical semantic decoding after the engineering receiver decodes the signal back into a message (e.g., the pulses of Morse code into English letters). This "semantic receiver" – currently just a black box – would match the statistical semantic characteristics of the message to the statistical semantic characteristics of the totality of receivers or the subset of them that the source wishes to affect. Within this black box, the causal relations of informational semantics would appear as statistical or probabilistic patterns of agent-world interaction. From this perspective, the man in Searle's (op.cit.) Chinese room does not understand the symbols he manipulates because his rulebook only embodies, metaphorically, the engineering receiver.

II. 5. 1982: Marr: Explanation

While many of the elements in the core explanatory package were discovered or derived from work that occurred during World War II, the post-war period involved the institutionalization of cognitive science and the development of these ideas within recognized institutional and disciplinary strictures. (Sept. 11, 1956 – the second day of a three-day Symposium on Information Theory at MIT – has been cited (Miller 2003: 142; Bechtel et al. op.cit.: 37) as an unofficial birthdate of cognitive science, but nothing hangs on this date.) Marr, a vision scientist, drew some general explanatory lessons from the emerging information-processing framework. Reacting to his contemporaries' focus on the physiology of single neurons in visual processing, Marr held that a full explanation of vision would require understanding not just physical mechanisms but also their organization and contexts of operation.

Marr (1982: 24) proposed that explaining any information-processing system required answering three different sorts of questions about it. These could be described

[10] Note that Dretske's (1981, 1983) appropriation of Shannon amounted to a causal theory of content of individual thoughts; as Dretske himself admits (1983: 82), he took very little from Shannon's actual theory.

and conceptualized in terms of three explanatory levels or analyses (Bechtel and Shagrir 2015; Shagrir 2010). The computational level involved explaining the why or goal of a particular kind of processing: What is the problem that the system need to solve? The algorithmic level involved explaining how this goal could be achieved in terms of the steps or state transitions leading to it: What sorts of representations and rules are used to solve the problem? The implementation level involved explaining how physical structures might realize these state transitions: What physical mechanisms instantiated these representations and their processing? Marr's approach yielded a common explanatory currency for integrating cognitive science research across disciplines, from neurobiology to cognitive psychology.

Marr, with his collaborators Tomaso Poggio and Ellen Hildreth, illustrated this approach by reframing visual processing into the same classical computational terms that were being used to explain higher cognitive capacities. Information-processing was not just about playing chess, but also perceiving objects. Systems within human agents could also be understood in the same basic information-processing terms. For example, activity in a particular area of V1 was for edge detection (computational level). It achieved this goal using rules for calculating zero-crossings (algorithmic level); and neural and other biological and biochemical machinery in this area of V1 implemented these algorithms. V1 is the most common label for the tip of the occipital lobe, at the back of the brain, where visual information is initially processed after passing through the retinas and subcortical brain structures. Additional processing in other visual areas would eventually yield a 3D image of an object.

The three levels of analysis could apply to many complex systems. Answers to any one of questions would provide constraints on answers to the others. So explaining any one system would require referring to systems at other levels:

> It is a feature of such [complex information-processing] tasks, arising from the fact that the information processed in the machine is only loosely constrained by the physical properties of the machine, that they must be understood at different, though inter-related, levels. (1981: 258)

In the case of vision, without answers at all three levels, describing the activity of neurons in response to specific stimuli, and even how these neurons are connected, would not yield an explanation of the phenomenon of vision. The need for multiple explanatory levels is hardly limited to cognition (O'Malley et al. 2013).

Other than in machine vision, Marr's emphasis on finding algorithms by which visual-feature outputs are computed (as in edge detection) has been superceded by a greater focus on real-world perception and embodied cognition (e.g., O'Regan and Noe 2001) and neural network computing methods (e.g., Olshausen and Field 1996). Debate over the necessity for symbolic representations in cognitive science has also sparked debate regarding the necessity for the algorithmic level in particular, given the classical computational terms in which Marr stated his theory. The relative independence of the levels, or the answers to the questions, has also been a matter of debate (although Marr and Poggio emphasized their interdependence). More recently, the contemporary search for canonical neural computations (Carandini 2012) pushes the explanatory framework downwards, while Poggio (2012: 1021) pushes it outward by adding learning and development to the three original levels.

Part III. Conclusion: What Lies Ahead?

As long as the mind remains a black box, there will always be a donkey on which to pin dualist ... intuitions.

Greene and Cohen 2004: 1781

The first century of cognitive science was largely a matter of formulating the basic explanatory package for materialism and exploring how much could be explained by these ideas. Different disciplinary specialization interpreted that framework in the ways most suited to their available technologies, training, and immediate explanatory goals. We do not yet have a comprehensive materialism, but there are advances going on in every direction.

One important example may be theories of consciousness (Dehaene et al. 1998, Oizumi et al. 2014, Tononi and Koch 2014), which had largely been left to philosophers (e.g., Chalmers op.cit., Block 1995) during "a century of taboo" in science (Baars 2003: fn. 1). This acceptance of consciousness as a scientific explanandum has been accompanied by efforts to accept reports of introspectively accessible conscious states as valid evidence (Jack and Roepstorff 2002, 2003). Clinical cases (e.g., detection of neural activity in vegetative-state patients), research in animal cognition, and advances in robotics are contributing to this final rejection of radical behaviorism and dualism.

New discoveries in neuroscience are also altering traditional ways of thinking about the mind. For example, the perception/cognition distinction is under siege given the discovery of the huge cortical allocation in higher primates to visual processing and new theories of vision in which the goal of vision is recognizing meaningful social stimuli (Nakayama 2010: 15). In memory research, our intuitive concept of memory as something stored in the brain, rather than constructed and elaborated in context, seems to get human memory wrong and computer memory right. Neuroimaging studies show overlap in brain areas involved in remembering past experiences and imagining or simulating possible future experiences. This suggests that remembering and imagining may be forms of a single process for preparing for the future, rather than distinct processes of recalling a stored representation and engaging in stimulus-independent thought (Schachter et al. 2012).

While the 21st century has already been dubbed the century of the brain (Flavell 2000), it is also likely to be the century of the social (see also Bechtel et al. op.cit.: 90). The fact that early conceptual innovations regarding social cognition arose from field work with animals (e.g. Jolly op.cit.) may explain why they were not integrated earlier: the very idea of animal cognition was and to some degree remains a matter of debate (Shettleworth 2010). But enactivist and embodied cognitive research points in the opposite direction from that recommended in Fodor's (1980) brief for methodological solipsism, a pragmatic recommendation for research modeled on Descartes's solipsistic method for discovering the essence of mind. This push away from solipsism has been a thread within cognitive science for some time (e.g. Thelen and Smith 1994, Gibson 1979, Brooks 1990, 1991). Social context is now being theorized in terms of multi-agent systems engaged in cooperation, communication, and learning.

It seems likely that the basic conceptual package for explaining agency will soon be fully elaborated in outline if not in its empirical details. Near the start of the last century, psychologist Karl Lashley summed up the materialist viewpoint as follows:

> The vitalist cites particular phenomena … and denies the possibility of a mechanistic account of them. But he thereby commits what we might call the egotistic fallacy. On analysis, his argument reduces every time to the form, "*I* am not able to devise a machine that will do these things; therefore no one will ever conceive of such a machine. (1923: 269)

If one substitutes "dualism" for "vitalism", a similar remark might be made regarding cognitive science at the start of the 21st century. Dualism will always remain conceivable, but an empirically testable theoretical framework for materialism is just a matter of time.

<u>Bibliography (2,300 words)</u>

Adams, F. and K. Aizawa (2008). *The Bounds of Cognition*. Oxford, UK: Wiley-Blackwell.

Adriaans, P. (2012). Information. *The Stanford Encyclopedia of Philosophy* (Fall 2013 edition), E. Zalta, ed., URL = http://plato.stanford.edu/archives/fall2013/entries/information/

Aizawa, K. (1992). Connectionism and artificial intelligence: history and philosophical interpretation. *Journal of Experimental and Theoretical Artificial Intelligence* 4: 295-313.

Allen, C., G. Varner, and J. Zinser (2000). Prolegomena to Any Future Artificial Moral Agent. *Journal of Experimental and Theoretical Artificial Intelligence* 12 (3): 251-61.

Arbib, M. (2000). McCulloch's Search for the Logic of the Nervous System. *Perspectives in Biology and Medicine* 43 (2): 193-216.

Aspray, W. (1985). The Scientific Conceptualization of Information. *Annals of the History of Computing* 7 (2): 117-40.

beim Graben, P. and J. Wright (2011). From McCulloch-Pitts Neurons Toward Biology. *Bulletin of Mathematical Biology* 73: 261-5.

Baars, B. (2003). The Double Life of B.F. Skinner. *Journal of Consciousness Studies* 10 (1): 5-25.

Baddeley, A. (1994). The Magical Number Seven: Still magic after all these years?

Psychological Review 101 (2): 353-6.

Bainbridge, W., E. Brent, K. Carley, D. Heise, M. Macy, B. Markovsky, and J. Skvoretz (1994). Artificial social intelligence. *Annual Review of Sociology* 20: 407-36.

Baronchelli, A., R. Ferrer-i-Cancho, R. Pastor-Satorras, N. Chater, and M. Christiansen (2013). Networks in Cognitive Science. *Trends in Cognitive Sciences* 17 (7): 348-60.

Bechtel, W., Abrahamsen, A., & Graham, G. (1998). The life of cognitive science. In W. Bechtel, & G. Graham, eds., *A Companion to Cognitive Science* (Oxford: Basil Blackwell: 1–104.

Bechtel, W. (2009). Constructing a Philosophy of Science of Cognitive Science. *Topics in Cognitive Science* 1: 548-69.

Bechtel, W. and O. Shagrir (2015). The Non-Redundant Contributions of Marr's Three Levels of Analysis for Explaining Information-Processing Mechanisms. *Topics in Cognitive Science* 7: 312-22.

Bender, A., E. Hutchins, and D. Medin (2010). Anthropology in Cognitive Science. *Topics in Cognitive Science* 2: 374-85.

Bender, A., S. Beller, and D. Medin (2012). Turning Tides: prospects for more diversity in cognitive science. *Topics in Cognitive Science* 4: 462-6.

Block, N. (1995). On a Confusion about a Function of Consciousness. *Behavioral and Brain Sciences* 18: 227-287.

Block, N. (2007). Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behavioral and Brain Sciences* 30 (5-6): 481-99.

Boden, M. (1991). Horses of a different color? In W. Ramsey, S. Stich, and D. Rumelhart, eds., *Philosophy and Connectionist Theory*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Boden, M. (2006). *Mind As Machine: a history of cognitive science*. Oxford: Oxford University Press.

Brooks, R. (1990). Elephants Don't Play Chess. *Robotics and Autonomous Systems* 6 (1-2): 3-15.

Brooks, R. (1991). Intelligence Without Representation. *Artificial Intelligence* 47: 139-59.

Byrne, W. and A. Whiten 1988. *Machiavellian Intelligence: social expertise and the evolution of intellect in monkeys, apes, and humans*. Oxford: Oxford University

Press.

Calvo, P. and F. Keijzer (2009). Cognition in Plants. In F. Balušcka, ed., *Plant-Environment Interactions: signaling and communication in plants*. Berlin and Heidelberg: Springer-Verlag.

Carandini, M. (2012). From circuits to behavior: a bridge too far? *Nature Neuroscience* 15 (4): 507-09.

Chomsky, N. (1959). A Review of B.F. Skinner's *Verbal Behavior*. *Language* 35 (1): 26-58.

Chalmers, D. (1995). *The Conscious Mind*. Oxford: Oxford University Press.

Chemero, A. (2011). *Radical Embodied Cognitive Science*. Cambridge, MA: MIT Press.

Churchland, P. M. (1990). On the Nature of Theories: a neurocomputational perspective. In W. Savage, ed., *Scientific Theories* (Minnesota Studies in the Philosophy of Science vol. 14): 59-101.

Churchland, P. S. (1986). *Neurophilosophy: Towards a unified science of the mind-brain*. Cambridge: MIT Press.

Churchland, P.S., and T. Sejnowski (1992). *The Computational Brain*. Cambridge: MIT Press.

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

Clark, A. (2013). Whatever Next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences* 36 (3): 181-204.

Copeland, B. J. (2008). The Modern History of Computing. *The Stanford Encyclopedia of Philosophy* (Fall 2008 edition), E. Zalta, ed., URL = http://plato.stanford.edu/archives/fall2008/entries/computing-history/

Copeland, B. J. and O. Shagrir (2013). Turing versus Godel on computability and the mind. In Copeland, B.J., C. Posy, and O. Shagrir, eds., *Computability: Turing, Godel, Church, and beyond*. MIT Press.

Dayan, P. and L. Abbott (2001). *Theoretical Neuroscience: computational and mathematical modeling of neural systems.* Cambridge, MA: MIT Press.

Dayan, P. (2012). Twenty-five Lessons From Computational Neuromodulation. *Neuron* 76: 240-56.

Dehaene, S., M. Kerszberg, and J.-P. Changeux (1998). A Neuronal Model of a Global Workspace in Effortful Cognitive Tasks. *Proceedings of the National Academy of Sciences USA* 95 (24): 14529-34.

Dennett, D. (1975). Brain Writing and Mind Reading. In K. Gunderson, ed., *Language, Mind, and Knowledge* (Minneapolis: University of Minnesota Press): 403-15.

Dennett, D. (1987). *The Intentional Stance*. Cambridge, MA: MIT Press.

Dretske, F. (1981). *Knowledge and the Flow of Information*. Cambridge, MA: MIT Press.

Dretske, F. (1983). Precis of Knowledge and the Flow of Information. *Behavioral and Brain Sciences* 6: 55-90.

Dretske, F. (1988). *Explaining Behavior*. Cambridge, MA: MIT Press.

Dreyfus, H. (1992). *What Computers Still Can't Do: a critique of artificial reason.* Cambridge, MA: MIT Press.

Dunbar, R. (1998). The Social Brain Hypothesis. *Evolutionary Anthropology* 6: 178-90.

Eliasmith, C. (2013). *How to Build a Brain: a neural architecture for biological cognition.* New York and Oxford: Oxford University Press.

Emery, N., N. Clayton, and C. Frith (2006). Introduction: Social intelligence: from brain to culture. *Philosophical Transactions of the Royal Society B* 362: 485-8.

Emery, N. and N. Clayton (2004). The Mentality of Crows: convergent evolution of intelligence in corvids and apes. *Science* 306 (5703): 1903-7.

Fodor, J. (1975). *The Language of Thought*. Thomas Y. Crowell.

Fodor, J. (1980). Methodological solipsism considered as a research strategy in cognitive psychology. *Behavioral and Brain Sciences* 63: 63-73.

Fodor, J. (1983). *The Modularity of Mind*. Cambridge, MA: MIT Press.

Fodor, J. and Z. Pylyshyn (1988). Connectionism and cognitive architecture: a critical analysis. *Cognition* 28: 3-71.

Froese, T., C. Gershenson, and L. Manzanilla (2014). Can government be self-organized? A mathematical model of the collective social organization of ancient Teotihuacan, central Mexico. *PLoS One* 9 (10), e109966.

Gallagher, S. (2005). *How the Body Shapes the Mind*. New York: Oxford University Press.

Gardner, H. (1985). *The Mind's New Science*. New York: Basic Books.

Gibson, J.J. (1979). *The Ecological Approach to Visual Perception*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Godfrey, M. and D. Hendry (1993). The Computer as von Neumann Planned It. *IEEE Annals of the History of Computing* 15 (1): 11-21.

Greene, J. and J. Cohen (2004). For the Law, Neuroscience Changes Everything and Nothing. *Philosophical Transactions: Biological Sciences* 359 (1451): 1775-85.

Greenwood, J. (1999). Understanding the "cognitive revolution" in psychology. *Journal of the History of the Behavioral Sciences* 35 (1): 1-22.

Grice, H. P. (1957). Meaning. *Philosophical Review* 66 (3): 377-88.

Guizzo, E. 2003 The Essential Message: Shannon and the Making of Information Theory (MIT: Masters' thesis in science writing). http://dspace.mit.edu/bitstream/handle/1721.1/39429/54526133.pdf;jsessionid

Hartley, R. (1928). Transmission of information. *Bell System Technical Journal* 7 (3): 535-63.

Hebb, D. (1949). *The Organization of Behavior*. New York: Wiley.

Heckathorn, D. (1989). Cognitive Science, Sociology, and the Theoretic Analysis of Complex Systems. *Journal of Mathematical Sociology* 14 (2-3): 97-110.

Hohwy, J. (2014). *The Predictive Mind*. New York: Oxford University Press.

Hopfield, J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the USA* 79 (8): 2554-58.

Humphrey, N. (1976). The social function of intellect. In P. Bateson and R. Hinde, eds., *Growing Points in Ethology* (Cambridge: Cambridge University Press): 303-17.

Izhikevich, E. (2007). *Dynamical Systems in Neuroscience: the geometry of excitability and bursting.* Cambridge, MA: MIT Press.

Jack, A. and A. Roepstorff (2002). Introspection and Cognitive Brain Mapping: from stimulus-response to script-report. *Trends in Cognitive Sciences* 6 (8): 333-9.

Jack, A. and A. Roepstorff (2003). Why Trust the Subject? Editorial Introduction to Trusting the Subject? The use of introspective evidence in cognitive science.

Journal of Consciousness Studies 10 (9-10): v-xx.

Jolly, A. (1966). Lemur Social Behavior and Primate Intelligence. *Science* (New Series), 153 (3735): 501-6.

Kiverstein, J. and M. Miller (2015). The embodied brain: towards a radical embodied cognitive neuroscience. *Frontiers in Human Neuroscience* (9), article 237, doi: 10.3389/fnhum.2015.00237

Lashley, K. (1923). The behavioristic interpretation of consciousness I and II. *Psychological Review* 30 (4): 237-72 and 30 (5): 329-53.

Lettvin, J., H. Maturana, W. McCulloch, and W. Pitts (1959). What the Frog's Eye Tells the Frog's Brain. *Proceedings of the IRE* 47: 1940-51.

Lewis, D. (1969). *Convention*. Cambridge: Harvard University Press.

Lombardi, O. (2005). Dretske, Shannon's Theory and the Interpretation of Information. *Synthese* 144 (1): 23-39.

Marcus, G. (2003). *The Algebraic Mind: integrating connectionism and cognitive science.* Cambridge, MA: MIT Press.

Marr, D. (1982). *Vision: a computational investigation into the human representation and processing of visual information*. San Francisco, CA: W.H. Freeman and Company.

Marr, D. and E. Hildreth (1980). Theory of Edge Detection. Proceedings of the Royal Society of London, Series B, Biological Sciences 207 (1167): 187-217.

Marr, D. and T. Poggio (1977). A Computational Theory of Human Stereo Vision. Proceedings of the Royal Society of London, Series B, Biological Sciences 204 (1156): 310-28.

McCulloch, W. and W. Pitts (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics* 5: 115-33.

Miller, G., E. Galanter, and K. Pribram (1960). *Plans and the Structure of Behavior*. New York: Holt, Rinehart & Winston.

Miller, G. (2003). The cognitive revolution: a historical perspective. *Trends in Cognitive Sciences* 7 (3): 141-4.

Miller, G. (1956). The Magical Number Seven, Plus or Minus Two. *Psychological Review* 63: 81-97.

Millikan, R. (1984). *Language, Thought, and other Biological Categories.* Cambridge, MA: MIT Press.

Minsky, M. and S. Papert (1969). *Perceptrons*. Cambridge, MA: MIT Press

Nakayama, K. (2010). Introduction: Vision Going Social. In R. Adams, Jr., N. Ambady, K. Nakayama, and S. Shimojo, eds., *The Science of Social Vision* (New York and Oxford: Oxford University Press): 3-17.

Nature (1948). Calculating Machines (unsigned summary of Royal Society discussion of March 4, 1948). *Nature* 161 (May 8): 712-13.

Neisser, U. (1967). *Cognitive Psychology.* New York: Meredith Publishing.

Newell, A. (1982). The Knowledge Level. *Artificial Intelligence* 18: 87-127.

Newell, A. (1993). Reflections on the Knowledge Level. *Artificial Intelligence* 59: 31-8.

Newell, A., J. Shaw, and H. Simon (1958). Elements of a Theory of Human Problem Solving. *Psychological Review* 65 (3): 151-66.

Newell, A. and H. Simon (1972). *Human Problem-Solving*. Upper Saddle River, NJ: Prentice-Hall.

Newell, A. and H. Simon (1976). Computer Science as Empirical Inquiry: Symbols and search. *Communications of the ACM (Association for Computing Machinery)* 19 (3): 113-126.

Newell, A. (1988). The Intentional Stance and the Knowledge Level. *Behavioral and Brain Sciences* 11 (3): 520-22.

Nyquist, H. (1924). Certain factors affecting telegraph speed. Bell System Technical Journal.

Oizumi, M., L. Albantakis, and G. Tononi (2014). From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. *PLoS Computational Biology* 10 (5), e1003588.

Olshausen, B. and D. Field (1996). Emergence of Simple-Cell Receptive Field Properties by Learning a Sparse Code for Natural Images. *Nature* 381: 607-09.

O'Malley, M., I. Brigandt, A. Love, J. Crawford, J. Gilbert, R. Knight, S. Mitchell, and F. Rohwer. Multilevel Research Strategies and Biological Systems. *Philosophy of Science* 81 (5): 811-28.

O'Regan, J.K. and A. Noe (2001). A Sensorimotor Account of Vision and Visual Consciousness. Behavioral and Brain Sciences 24: 939-1031.

Peckhaus, V. (2009). Leibniz's Influence on 19th Century Logic. *The Stanford Encyclopedia of Philosophy* (Spring 2014 edition), E. Zalta, ed., URL = http://plato.stanford.edu/archives/spr2014/entries/leibniz-logic-influence/

Piccinini, G. (2004). The first computational theory of the brain: a close look at McCulloch and Pitts's "Logical Calculus of Ideas Immanent in Nervous Activity". *Synthese* 141 (2): 175-215.

Piccinini, G. and A. Scarantino (2010). Computation vs. information processing: why their difference matters to cognitive science. *Studies in History and Philosophy of Science* 41: 237-46.

Poggio, T. (1981). Marr's computational approach to vision. *Trends in Neurosciences* 10: 258-62.

Poggio, T. (2012). The Levels of Understanding Framework, revised. *Perception* 41: 1017-23.

Pospichal, J. and V. Kvasnicka (2015). 70th Anniversary of Publication: Warren McCulloch and Walter Pitts – A Logical Calculus of the Ideas Immanent in Nervous Activity. In P. Sincak et al., eds., *Emergent Trends in Robotics and Intelligent Systems*. Switzerland: Springer International Publishing.

Rogers, T. and J. McLelland (2014). Parallel Distributed Processing at 25: further explorations in the microstructure of cognition. *Cognitive Science* 38: 1024-77.

Rosenblatt, F. (1958). The Perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review* 65: 386-408.

Rosenblueth, A., N. Wiener, and J. Bigelow (1943). Behavior, Purpose, and Teleology. *Philosophy of Science* 10 (1): 18-24.

Roskies, A. (2010). How Does Neuroscience Affect Our Conception of Volition? *Annual Reviews Neuroscience* 33: 109-30.

Rumelhart, D., J. McClelland, and the PDP Research Group (1986). *Parallel Distributed Processing: Explorations in the microstructure of cognition.* Vol. 1: Foundations, and Vol.2: Psychological and biological models. Cambridge, MA: MIT Press.

Schacter, D., D. Addis, D. Hassabis, V. Martin, R. N. Spreng, and K. Szpunar (2012). The Future of Memory: Remembering, imagining, and the brain. *Neuron* 76 (4): 677-94.

Searle, J. (1980). Minds, Brains, and Programs. *Behavioral and Brain Sciences* 3: 417-24.

Shagrir, O. (2010). Marr on Computational-Level Theories. *Philosophy of Science* 77

(4): 477-500.

Shannon, C. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal* 379-423, 623-56.

Simon, H. (1996). *The Sciences of the Artificial* (3rd ed.). Cambridge, MA: MIT Press.

Simon, H. (1974). How big is a chunk? *Science* 183: 482-88.

Smolensky, P. (1991). Connectionism, constituency and the language of thought. In B. Loewer and G. Rey, eds., *Meaning in Mind: Fodor and His Critics* (Oxford: Blackwell).

Sterelny, K. (2007). Social intelligence, human intelligence, and niche construction. *Philosophical Transactions of the Royal Society B* 362: 719-30.

Smith, E. (1996). What do Connectionism and Social Psychology Offer Each Other? *Journal of Personality and Social Psychology* 70 (5): 893-912.

Sutton, R. and A. Barto (1981). Towards a Modern Theory of Adaptive Networks: Expectation and Prediction. *Psychological Review* 88 (2): 135-70.

Sutton, R. and A. Barto (1998). *Reinforcement Learning: an introduction*. Cambridge, MA and London: MIT Press. Available online at: http://webdocs.cs.ualberta.ca/~sutton/book/ebook/the-book.html

Thelen, E. and L. Smith (1994). *A Dynamical Systems Approach to Development of Cognition and Action.* Cambridge, MA: MIT Press.

Tononi, G. and C. Koch (2014). Consciousness: Here, There, but Not Everywhere. arXiv preprint arXiv:1405.7089

Trewavas, A. (2005). Green plants as intelligent organisms. *Trends in Plant Sciences* 10: 413-19.

Turing, A. (1936). On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society* Series 2, 47: 230-65.

Turing, A. (1950). Computing Machinery and Intelligence. *Mind* 59 (236): 433-60.

van Gelder, Tim (1995). What Might Cognition Be, if Not Computation? *Journal of Philosophy* 92 (7): 345-81.

von Eckardt, B. (2003). The Explanatory Need for Mental Representations in Cognitive Science. *Mind & Language* 18 (4): 427-39.

von Neumann, J. (1945). First Draft of a Report on the EDVAC. Moore School of Electrical Engineering, University of Pennsylvania. Reset typescript online at: http://www.virtualtravelog.net/wp/wp-content/media/2003-08-TheFirstDraft.pdf

von Neumann, J. (1966). *Theory of Self-Reproducing Automata*. Edited and completed by Arthur W. Burks. Urbana and London: University of Indiana Press.

Weaver, W. (1949). The mathematics of communication. *Scientific American* 181 (1): 11-15.

Wiener, N. (1948). *Cybernetics, or Control and Communication in the Animal and the Machine.* New York: John Wiley & Sons.

Williams, F. and T. Kilburn (1948). Electronic Digital Computers. *Nature* 162 (4117): 487.